Some New Results on the Estimation of Sinusoids in Noise

Nielsen, Jesper Kjær

Publication date:
2012

Document Version
Accepted author manuscript, peer reviewed version

Some New Results on the Estimation of Sinusoids in Noise

Ph.D. Dissertation

Jesper Kjær Nielsen

Aalborg University
Department of Electronic Systems
Fredrik Bajers Vej 7B
DK-9220 Aalborg

Nielsen, Jesper Kjær
Some New Results on the Estimation of Sinusoids in Noise
ISBN 978-87-92328-91-5

Department of Electronic Systems
Aalborg University
Fredrik Bajers Vej 7
DK-9220 Aalborg Ø
Denmark

This thesis has been printed with Computer Modern 10pt and been typeset using LaTeX 2_ε on a computer running the GNU/Linux operating system. All of the figures have been created using GNUPLOT, PGF and the macro packages Ti*k*Z and PGFPLOTS. Simulations have been run in Matlab™.

Abstract

This thesis is concerned with the problem of estimating sinusoidal parameters from noisy observations. This field of research is applicable to solving problems in a large number of areas such as music and speech processing, electrocardiography, seismology, radar and sonar processing, astronomy, meteorology, and economics, and in this thesis a number of rather diverse contributions are made to this field of research. These contributions include new results and algorithms in relation to model comparison and selection, fundamental frequency estimation, inference in dynamic sinusoidal models, and filtering methods.

In the introductory part of this thesis, an overview over the modelling and inference problem is given, and the most important methods for solving these problems are briefly reviewed. During this introduction, the contributions are also stated and positioned in relation to these previously proposed methods. The second part of this thesis contains the contributions. First, the model comparison and selection problem is considered for a general non-linear model. In this connection, a few new model comparison methods are proposed and demonstrated to perform better than existing methods for both model selection and prediction. Second, the joint fundamental frequency estimation and model order detection problem is analysed within a Bayesian framework. A new method is also suggested and its accuracy is evaluated and demonstrated to perform better than a similar state-of-the-art method. Third, an efficient algorithm for performing inference and interpolation in a dynamic sinusoidal model is proposed. This method is applied to packet-loss concealment, and listening tests indicate that the proposed algorithm can be used for this purpose. Fourth, the Capon filtering method for amplitude estimation is extended in an interesting way by selecting the filter length of the Capon filter in a data-adaptive fashion. Finally, the recently proposed sampling scheme called compressed sensing is analysed in the context of estimating continuous parameter such as the frequency parameter and the direction-of-arrival, and it is shown that compressed sensing decreases the estimation accuracy of such parameters.

Although the estimation problems considered in this thesis are primarily analysed in the context of speech and audio applications, the results are useful in a wider range of applications. Along these lines, the main focus has not been on developing new algorithms for specific applications, but rather on understanding the underlying estimation problem and analysing it in a consistent fashion.

Resumé

Nærværende afhandling omhandler forskellige problemstillinger i forbindelse med estimering af sinusparametre ud fra støjfyldte observationer. Dette forskningsområde er anvendeligt i forbindelse med at løse en række problemer inden for mange områder som for eksempel musik- og taleprocessering, elektrokardiografi, seismologi, radar- og sonarprocessering, astronomi, metrologi og økonomi, og en række forskelligartede bidrag til dette område beskrives i denne afhandling. Disse bidrag omfatter nye resultater og algoritmer i relation til sammenligning af modeller, estimering af grundfrekvensen, analyse og syntese med dynamiske sinusmodeller, og filtreringsmetoder.

I afhandlings introducerende del gives et overblik over estimerings- og modelleringsproblemet, og de vigtigste metoder til at løse disse problemer beskrives kortfattet. De videnskabelige bidrag, som beskrives i afhandlingens anden del, positioneres i forhold til de eksisterende metoder undervejs i introduktionen. Afhandlingens anden del indeholder de videnskabelige bidrag. I det første bidrag udvikles nogle nye metoder til at sammenligne forskellige modeller og udvælge den bedste. Disse metoder sammenlignes med tilsvarende metoder, og det vises, at de foreslåede metoder giver bedre resultater. Ved hjælp af bayesianske metoder, analyseres dernæst i det andet bidrag den samtidige estimering og detektion af et periodisk signals grundfrekvens og modelorden. En ny metode foreslås også og dens nøjagtighed evalueres. Det vises også, at den foreslåede metode forbedrer en tilsvarende og førende metode. I det tredje bidrag udvikles en effektiv algoritme til at foretage analyse og syntese af signaler ved hjælp af en dynamisk sinusmodel. Algoritmen bruges til at rekonstruerede tabte lydpakker i et pakkebaseret netværk, og lyttetests indikerer, at algoritmen kan bruges til dette formål. I det fjerde bidrag udbygges Capon filtreringsmetoden på interessant vis ved at vælge filterets længde på en selvjusterende måde. I det sidste bidrag analyseres den nye samplingsteknik kaldet compressed sensing i forbindelse med estimering af kontinuerte parametre som frekvensen og ankomstvinklen, og det vises, at compressed sensing forværrer estimeringsnøjagtigheden for disse parametre.

Selvom de estimeringsproblemer, der er behandlet i denne afhandling, primært er analyseret i forbindelse med tale- og musikapplikationer, kan de præsenterede resultater også bruges i mange andre sammenhænge. Det primære fokusområde har ikke været på at udvikle én ny algoritme til én bestemt applikation, men snarere at forstå det underliggende estimeringsproblem og analysere det på en konsistent måde.

Contents

List of Papers

The main body of this thesis consist of the following papers.

[A] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, and S. H. Jensen, "Bayesian model comparison with the g-prior," Submitted to *IEEE Trans. Signal Process.*.

[B] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "Default Bayesian estimation of the fundamental frequency," Submitted to *IEEE Trans. Audio, Speech, Lang. Process.*.

[C] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, S. J. Godsill, and S. H. Jensen, "Bayesian interpolation and parameter estimation in a dynamic sinusoidal model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1986–1998, Sep. 2011.

[D] ——, "Bayesian interpolation in a dynamic sinusoidal model with application to packet-loss concealment," in *Proc. European Signal Processing Conf.*, Aug. 2010.

[E] J. K. Nielsen, P. Smaragdis, M. G. Christensen, and S. H. Jensen, "An amplitude spectral Capon estimator with a variable filter length," to appear in *Proc. European Signal Processing Conf.*, Aug. 2012.

[F] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "On compressed sensing and the estimation of continuous parameters from noisy observations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012.

[G] M. G. Christensen and J. K. Nielsen, "Joint direction-of-arrival and order estimation in compressed sensing using angles between subspaces," in *Proc. IEEE Workshop on Stat. Signal Process.*, Jun. 2011, pp. 449–452.

All documents and the code used in the simulation of the published documents are available at `http://kom.aau.dk/~jkn/publications/publications.php`. In addition to the main papers, the following publications have also been made.

[1] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "An approximate Bayesian fundamental frequency estimator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012.

[2] J. K. Nielsen and S. H. Jensen, "Lecture notes in adaptive filters," Feb. 2012, Aalborg University. [Online]. Available: http://kom.aau.dk/~jkn/publications/publications.php

[3] J. K. Nielsen, J. R. Jensen, M. G. Christensen, S. H. Jensen, and T. Larsen, "Waveform approximating residual audio coding with perceptual pre- and post-filtering," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Oct. 2008.

[4] J. R. Jensen, J. K. Nielsen, M. G. Christensen, S. H. Jensen, and T. Larsen, "On fast implementation of harmonic MUSIC for known and unknown model orders," in *Proc. European Signal Processing Conf.*, Aug. 2008.

I should also state that although not of the Bayesian persuasion, I am willing to consider any technique, regardless of its motivation.

Anonymous reviewer

Preface

This thesis is submitted to the International Doctoral School of Technology and Science at Aalborg University in a partial fulfilment of the requirements for the degree of Doctor of Philosophy. The work was carried out in the period spanning from August 2009 to July 2012 at the Department of Electronic Systems at Aalborg University.

The thesis is concerned with the estimation of sinusoidal parameters from noisy observations and is divided into two parts. In the first part, an overview is given to the estimation problem and previously proposed methods are reviewed. The main body of the thesis is its second part. This part consists of a number of papers which have been published in or submitted to peer-reviewed conferences or journals. As the range of contributions is rather diverse, the papers have been ordered according to both their significance and their similarity.

I would like to thank my main supervisors Prof. Søren Holdt Jensen and Assoc. Prof. Mads Græsbøll Christensen for providing me with the freedom to pursue my interests, for giving me helpful guidance, and for supporting me. They have been an integral part of the results that I have achieved, and they deserve my gratitude. They have also been a major part in setting up my three months visit to Ass. Prof. Paris Smaragdis at University of Illinois at Urbana-Champaign in October-December 2011. In that connection, I would also like to thank Paris Smaragdis for being an inspirational source of information and for spending some of his precious time on supervising me. I would also like to thank Prof. Simon Godsill, University of Cambridge, and Ass. Prof. Ali Taylan Cemgil, Boğaziçi University, for their collaboration, fruitful discussions, and for reading and answering my numerous e-mails.

The other Ph.D.-students here at Aalborg University also deserve to be acknowledged for their technical and/or social contributions to my life in the past three years. Especially, my office-buddy Jesper Rindom Jensen deserves credit for our close collaboration and fruitful discussions in the past seven years where he has also been an integral part of my social life. Last, but not least, I would also like to thank my friends and family for their love and support.

Jesper Kjær Nielsen
Aalborg University, July 4, 2012

Part I

Introduction

Sinusoids in Noise

1 Periodic Signals

Many natural and artificial phenomena exhibit some kind of cyclical behaviour. For example, the human heart contracts and relaxes once per cardiac cycle, the sun rises and sets once per day, crops are planted and harvested once per cropping season, the temperature varies on a daily and yearly basis, the sunspot activity peaks once per solar cycle, and the regent of Denmark gives a New Year's speech once a year. Knowing the duration of a cycle, also simply called the period, is therefore vital in order to be able to make predictions about the future or to facilitate a better understanding of the observed phenomena. Often, however, the period is unknown or varying with time, and it must therefore be estimated from the available data. In Fig. 1, an example of a famous data set known as the Wolf's relative sunspot numbers is shown. The data set is named after Rudolf Wolf who formalised the counting of the number of sunspots in 1848 and collected earlier scattered observations dated as far back as to the beginning of the 17th century [118]. From Fig. 1, we see that the number of sunspots exhibits a cyclical behaviour with a period of approximately 11 years. Estimating this period can be useful to understand some of the physical processes in the sun and to predict the Earth's climate, the financial periods, and the electromagnetic communication conditions in the ionosphere [209]. Another data set originating from a completely different source is shown in Fig. 2. The figure displays a segment of recorded female speech, and it clearly reveals that the speech waveform or signal also exhibits a cyclical behaviour. Estimating the period of the speech signal is useful in applications such as speaker identification [11, 135], automatic speech recognition [78, 95, 129], speech coding and compression [78, 137, 207, 227], speech enhancement [78, 137, 156, 227], and speech separation [158, 194]. A plethora of other applications exists in which the cycle periods must be estimated, and the estimation of these periods has therefore been subject to extensive research for several decades [203]. In this thesis, some new results are presented in relation to this estimation problem. Although the context is mainly that of speech and audio applications, the methods are applicable to solving problems in other fields as well. The first part of this thesis is concerned with a brief overview over some of

the central sinusoidal models and existing estimation methods. In this connection, we discuss their strengths and weaknesses, and list our contributions to the field. These contributions are described in much greater detail in a number of papers constituting the second part of the thesis.

1.1 Spectral Estimation

Due to the large scale applicability of finding cycle periods in data sets, the scientific field of spectral estimation has emerged. In their book on spectral estimation, Stoica and Moses defined the spectral estimation problem in the following way [215, p. 1].

> From a finite record of a stationary data sequence, estimate how the total power is distributed over frequency.

The frequency is the number of cycles per second and is thus the inverse of the cycle period. The term *stationary* describes a technical requirement on the statistical properties of the underlying process or signal from which a data set is observed. Specifically, stationary should here be interpreted in the weak- or wide-sense, meaning that the mean and the covariance functions of the signal must be time-invariant. Although many signals are nonstationary, they are usually approximately stationary on a local scale [131, p. 4]. A speech signal is an example of a nonstationary signal which may be approximately wide-sense stationary (WSS) on a local scale as shown in Fig. 2.

The plot of the power as a function of the frequency is called the power spectral density (PSD), and in Fig. 3, the simple periodogram estimate of the PSD is shown for the speech segment in Fig. 2. The periodogram reveals two things. First, there are two dominating positive cycles in the speech segment at approximately 230 Hz and 460 Hz. Second, the frequencies of the dominating and the inferior cycles seem to be an integer multiple of the longest cycle with the frequency of approximately 230 Hz. A signal with this special structure in the frequency domain is called a periodic signal, and it has the property that it exactly repeats itself for a time-shift equal to the longest cycle period.

1.2 Parametric Modelling

The periodogram is an example of a non-parametric approach to spectral estimation. The non-parametric methods have the advantage that they do not assume anything but wide-sense stationarity about the signal under study. However, the lack of the model assumption means that the non-parametric methods are required to estimate the infinite number of points constituting the PSD from a finite number of data point, and this leads to a large variance or a poor resolution of the non-parametric methods [215, pp. 217–218]. On the other hand, parametric approaches assume a parametric model for the signal under study, and the spectral estimation problem therefore reduces to the problem of estimating a number of model parameters which is usually much smaller

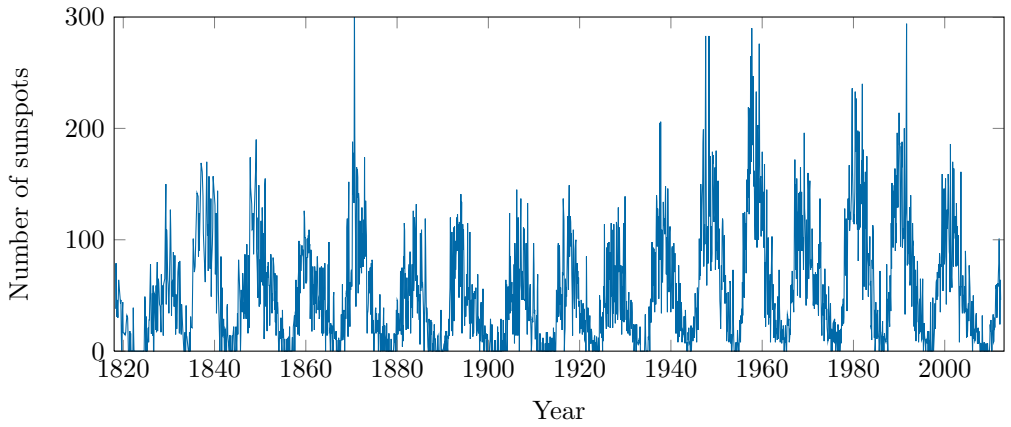

Fig. 1: The daily International Sunspot Number since January 1, 1818. The data are collected and provided by the Solar Influences Data Analysis Center in Belgium [205].

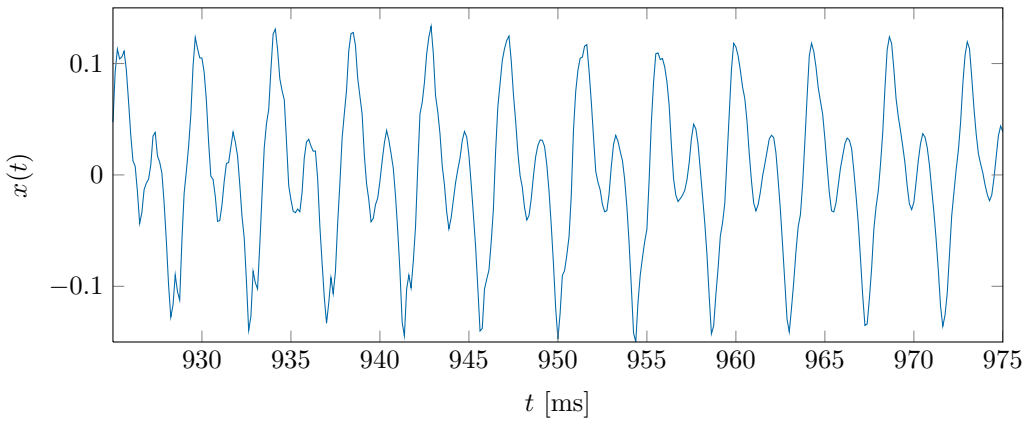

Fig. 2: A 50 ms segment from a recorded speech signal of a female uttering, 'Why where you away a year, Roy?'

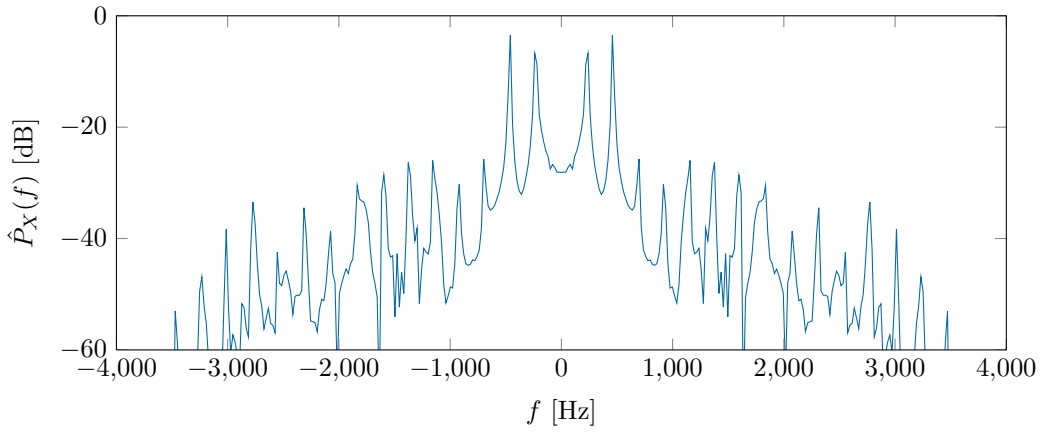

Fig. 3: The periodogram estimate of the PSD of the speech segment shown in Fig. 2.

than the number of data points. The most popular models are the autoregressive (AR) model, moving average model (MA), the autoregressive moving average model (ARMA), and the sinusoidal model, and spectral estimation using these models have received significant research attention (see, e.g., [25, 131, 215] for an overview). There exist many versions and combinations of the different models to account for physical phenomena such as amplitude and frequency modulation [52, 65, 143, 223, 224], transients [63, 105, 116, 127], and correlated noise [47, 75, 211]. Although many of these signal models model nonstationary phenomena which do not have a PSD, parameter estimation is still possible, and in this thesis, we focus on the estimation problem.

In relation to parametric modelling, the two data sets shown in Fig. 1 and Fig. 2 give rise to two important questions. First, the underlying generating process is presumably unknown which means that the true model is unknown. Since the data sets exhibit a cyclical behaviour, it seems reasonable to represent the generative process by a sinusoidal model, but how many cyclical components or sinusoids are there in the data? Second, what should be assumed about the non-cyclical part which is often referred to as the noise? Due to these two questions, inference in parametric models is not only a problem of estimating the unknown model parameters, but also a problem of determining the best model among a set of candidate models.

1.3 Speech and Music Applications

The methods presented in this thesis are applicable to solving problems in a large number of areas such as speech and music processing, electrocardiography, seismology, radar and sonar processing, astronomy, meteorology, and economics. However, we are

here primarily concerned with applications involving speech and music signals, and below a brief overview is given over a few of these applications.

Coding and compression: Speech and audio coding can be performed efficiently by estimating, quantising, and encoding the parameters of a parametric model rather than compressing the signal waveform directly.

Recognition and transcription: The parameters of a model can say something about the information conveyed by the signal. For example, a computer can use this information to automatically transcribe or classify a piece of music or to convert spoken words into text or commands.

Restoration and enhancement: Distorted or degraded signals can be enhanced by extracting the signal part of the observed data. For example, the perceptual artefacts caused by scratches on a CD, lost or corrupted packets on a packet-based network, or bad sectors on a storage medium can be reduced [103].

Separation: A mixture of signals can be approximately separated from its parametric representation. For example, the fundamental periods in a so-called multi-pitch signal [51] can be used to separate a mixture of two speakers or musical instruments.

Modification: Performing time-scale modifications of an audio signal becomes much easier when a parametric model is used.

A more thorough overview over these and other applications can be found in [51, Sec. 1.6].

Speech and audio signals are generated by complicated and non-linear physical processes. Although both speech models [207, Ch. 2] and instrument-specific models [91] have been constructed, effects such as reverberation, background noise and mixing make it almost impossible to construct a single signal model that can take all such phenomena into account. Moreover, very complicated models typically entail impractical estimators with a high computational complexity. Therefore, a trade-off must be made between the model complexity and algorithmic simplicity. In Sec. 2, we review some of the useful and popular sinusoidal models, and in Sec. 3, several methods for estimating the parameters of these models are discussed. A much more complete discussion on models and estimation methods can be found in [51].

2 Sinusoidal Models

A sinusoidal model is the most common way of modelling a signal exhibiting cyclical behaviour, and it can in a very general form be written as

$$x(t) = \sum_{i=1}^{l} A_i(t) \cos\big(\theta_i(t)\big) + e(t) \ . \tag{1}$$

As the notation suggests, the variable t usually represents time in this thesis, but depending on the application it may be any continuous (or discrete) quantity such as length, pressure, or temperature. The functions $A_i(t)$ and $\theta_i(t)$ are the amplitude and angle, respectively, of the i'th sinusoidal component whose phase ϕ_i and frequency $f_i(t)$ are defined as [21, 115]

$$\phi_i \triangleq \theta_i(0) \tag{2}$$

$$f_i(t) \triangleq \frac{1}{2\pi} \frac{d\theta_i(t)}{dt} \ . \tag{3}$$

The integer l is the total number of sinusoidal components in our model, and $e(t)$ denotes the non-sinusoidal part which is typically referred to as the noise. This noise term is usually modelled as a time-varying ARMA (p, q) model or a special case thereof. One such important special case is the ARMA $(0, 0)$ model which is simply white Gaussian noise (WGN).

2.1 Complex Sinusoidal Models

Although all physical signals are real-valued, it might be advantageous from an analytical, a notational, or an algorithmic point of view to work with the signals in their analytic form which is complex-valued [94]. Moreover, complex-valued signals can appear due to for example complex demodulation of a real-valued signal [215, Ch. 6]. For a real-valued continuous-time signal $x(t)$, its analytic signal $x_a(t)$ is defined as [51, App. A]

$$x_a(t) \triangleq x(t) + j\mathcal{H}(x(t)) \tag{4}$$

where $j = \sqrt{-1}$ is the imaginary unit and $\mathcal{H}(\cdot)$ denotes the Hilbert transform operator which is a convolution between the original signal and the function $1/(\pi t)$ [175]. The analytic signal $x_a(t)$ has two fundamental properties. First, its real part is the same as the original real-valued signal $x(t)$, and, second, the signal spectrum of the analytic signal is 0 at all negative frequencies, the same at $f = 0$ Hz, and twice that of $x(t)$ at the positive frequencies [163]. As illustrated in Fig. 3, the negative side of the signal spectrum contains redundant information for a real-valued signal and can therefore be removed. The analytic signal is only defined for a continuous-time signal, but Marple derived in [163] a

discrete-time "analytic" signal with many of the same properties as the continuous-time analytic signal. Moreover, Marple also showed that the sampling rate can be reduced by a factor of two when working with the discrete-time "analytic" signal since the redundancy in the signal spectrum can be removed. This suggests that there might be an algorithmic advantage of working with the "analytic" signal instead of the real-valued signal although the algebra now involves complex-valued numbers [51, App. A]. However, it should be noted that if the original signal spectrum contains significant power at the frequencies close to 0 or the Nyquist frequency (relative to N), spectral leakage shifts the peaks of this signal spectrum. Moreover, the algorithm suggested in [163] for computing the discrete-time "analytic" signal may have poor characteristics such as excessive ripple between the frequency points of the signal [51, App. A]. Consequently, estimates based on the signal spectrum of the discrete-time "analytic" signal are biased, but for many practical applications the bias is insignificant [48], [51, App. A].

When we have either transformed a real-valued signal into its "analytic" form or are working directly with complex-valued data, the complex-valued pendant to the real-valued signal model in Eq. (1) is

$$x_a(t) = \sum_{i=1}^{l} A_i(t) \exp\left(j\theta_i(t)\right) + e_a(t) \ . \tag{5}$$

The noise term $e_a(t)$ is also complex-valued and can be modelled as a complex-valued ARMA process.

2.2 Special Cases

The real-valued and complex-valued sinusoidal models in Eq. (1) and Eq. (5), respectively, contain a lot of model parameters. Moreover, the numbers of sinusoids, autoregressive parameters, and moving average parameters are usually unknown and must therefore also be inferred from the data along with the model parameters. This is in general a very difficult task which typically require a lot of computational resources. Therefore, simpler special cases of Eq. (1) and Eq. (5) are usually considered instead. Besides the assumptions on the noise term $e(t)$, these special cases are different ways of selecting the amplitude function $A_i(t)$ and the angle function $\theta_i(t)$, and we here review some of the popular choices for these two functions. A more thorough review in the context of audio modelling can be found in [106].

The Basic Model

The most popular sinusoidal model is the basic model which is obtained by making a zeroth- and first-order Taylor expansion of $A_i(t)$ and $\theta_i(t)$, respectively, around $t = 0$

so that

$$A_i(t) = A_i \tag{6}$$
$$\theta_i(t) = 2\pi f_i t + \phi_i = \omega_i t + \phi_i \ . \tag{7}$$

where $\omega_i \triangleq 2\pi f_i$ is the angular frequency. This parametrisation allows the polar form of the real-valued model in Eq. (1) to be rewritten into a rectangular form as

$$x(t) = \sum_{i=1}^{l} \Big[a_i \cos\big(\omega_i t\big) - b_i \sin\big(\omega_i t\big) \Big] + e(t) \tag{8}$$

where $a_i \triangleq A_i \cos(\phi_i)$ and $b_i \triangleq A_i \sin(\phi_i)$ are the in-phase and quadrature components, respectively. A similar form exists for the complex-valued model in Eq. (5), and it is given by

$$x_a(t) = \sum_{i=1}^{l} \alpha_i \exp(j\omega_i t) + e(t) \tag{9}$$

where $\alpha_i \triangleq A_i \exp(j\phi_i) = a_i + jb_i$ is the complex amplitude. Although this model is the simplest sinusoidal model, it is used in a wide range of practical applications such as audio coding [59, 66, 90, 181, 228], speech coding [167, 168], packet-loss concealment [154, 155], and direction of arrival estimation [93, 139]. The basic model is also used in spectral estimation since it is WSS provided that the phase is assumed to be a uniform random variable on any continuous interval of length 2π [131, 215].

Harmonic Models

The harmonic models are closely related to the basic model. As many signals are approximately periodic, the frequency of the i'th sinusoidal component in Eq. (7) can be modelled as the function $\omega_i = h(i, \omega_0)$ where ω_0 is the fundamental frequency. For a perfectly periodic signal, the function is

$$h(i, \omega_0) = i\omega_0 \ , \tag{10}$$

and the sinusoidal components are called harmonics. Voiced speech and many musical instruments have nearly this harmonic structure, and in Fig. 4, a spectrogram of a voiced speech signal is shown. However, the frequencies of the harmonics in recordings from plucked string instruments and pianos do not obey the simple relationship in Eq. (10), and this is often referred to as inharmonicity. For stiff-stringed instruments, a more accurate model is [91, p. 64]

$$h(i, \omega_0) = i\omega_0 \sqrt{1 + Bi^2} \tag{11}$$

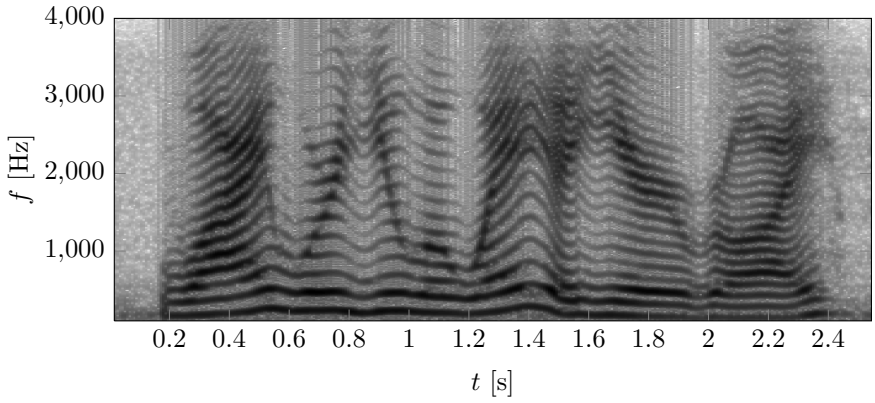

Fig. 4: The spectrogram of a female uttering, "Why were you away a year, Roy?". The spectrogram reveals a clear harmonic structure of the speech data.

where $B \ll 1$ is an instrument-specific stiffness parameter. The model for pianos is slightly more complex and can be found in [91, p. 363]. The instrument specific models have been used in a number of settings such as audio analysis and coding [64, 89, 102, 170, 236]. To avoid having an instrument specific model, the so-called perturbed model can be used as an alternative [51, 64, 74, 101]. It is given by

$$h(i, \omega_0) = i\omega_0 + \Delta_i \tag{12}$$

where Δ_i models the deviation from the harmonic grid of the perfectly periodic signal. Fundamental frequency estimation has received some attention particularly in the field of speech and audio processing [50, 51, 53, 75, 76, 130, 191, 197, 198, 220] where the fundamental frequency is often referred to as the pitch. Although the pitch refers to an auditory sensation rather than a physical attribute [179], the two terms are often used synonymously [51].

Models with Amplitude Modulation

For stationary signal segments, the constant amplitude assumption in the basic model is reasonable accurate. Hovewer, for nonstationary signal segments such as transient phenomena, the constant amplitude model is too simple. A remedy for this is to assume an amplitude function of the form

$$A_i(t) = \gamma_i(t) A_i \tag{13}$$

where $\gamma_i(t)$ is an envelope function. The envelope function can be selected in a very flexible way as a linear combination of basis functions [52], but the method suffers

from a high computational cost [225]. A popular alternative is an exponential damping function [27, 28, 105, 116, 127, 173] which involves a damping coefficient and an onset parameter. Unfortunately, the model is not smooth due to the discontinuity at the onset, and this may lead to some artefacts in for example audio processing where smoothness is a desirable feature [58, 63]. In [62, 63], the exponential damping function was therefore generalised to gamma functions which also model the smooth attack of the transients, thus avoiding the discontinuity of the exponential damping function.

In addition to the polar form of the sinusoidal model, amplitude modulation can also be modelled in the rectangular form of the sinusoidal model in Eq. (8) and Eq. (9). In for example [75, 89, 101], the in-phase and quadrature components were modelled as a linear combination of basis functions. Alternatively, these two components can be modelled as first-order Gauss-Markov processes [182, 223] as we have done in Paper C and Paper D where we have also shown that this is equivalent to a state space form of a so-called dynamic sinusoidal model [40–42]. The main advantages of this model are that it is physically more realistic [40] and makes the inference for the frequency parameter more tractable [223]. Within the field of econometrics, dynamic sinusoidal models are referred to as stochastic cyclical models, and they have also attracted some research attention [111–113, 136].

Models with Frequency Modulation

In audio signals, vibrato [2] and glissando [117] are examples of naturally occurring phenomena which can be modelled by frequency modulation. Other natural occurrences of frequency modulation are encountered in quantum optics, human speech, the navigational signals emitted by bats, and oceanography [124, 208]. Frequency modulation can be modelled by writing the angle function as [115, 161]

$$\theta_i(t) = \varpi_i \int_0^t q_i(\tau) d\tau + \omega_i t + \phi_i \tag{14}$$

for $|q_i(t)| \leq 1$ where ϖ_i is the maximum angular frequency deviation from ω_i, and $q_i(t)$ is the modulating signal. According to the definition of the frequency in Eq. (3), the frequency function is therefore given by

$$\omega_i(t) = \varpi_i q_i(t) + \omega_i \ , \tag{15}$$

and ϖ_i is usually much smaller than ω_i. Typical examples of the modulating signal $q_i(t)$ are a sine wave [166, 183], an autoregressive process [49, 223, 224], and linear functions (the so-called linear chirps or chirp signals) [17, 124, 166, 208].

Models for Sparse Decompositions

Estimating the nonlinear parameters such as the frequency or the stiffness parameter from a data set $\{x(t_n)\}_{n=0}^{N-1}$ might be difficult or computationally costly. To overcome

this issue, the sinusoidal models can be written as a sparse decomposition given by [43, 159]

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{s} + \boldsymbol{e} \tag{16}$$

where $\boldsymbol{x}$, $\boldsymbol{\Psi}$, $\boldsymbol{s}$, and $\boldsymbol{e}$ are the N-dimensional data vector, the $N \times D$-dimensional basis or dictionary, the D-dimensional S-sparse vector, and the N-dimensional noise vector, respectively. By S-sparse, we mean that $\boldsymbol{s}$ contains S non-zero components and $D - S$ zeros or, equivalently, that the noiseless part of $\boldsymbol{x}$ is represented by a weighted sum of S columns from the dictionary. The dictionary is assumed known and constructed from a physical signal model by sampling all but the linear parameters on a grid. As an example consider the model in Eq. (9). By sampling the frequency parameter on a uniform N-point grid, the dictionary is simply the Fourier basis [138]. Another very popular class of dictionaries are the wavelets bases [138, 159], but also a lot of other dictionaries such as chirplet and Gabor bases are used [43]. Whereas the Fourier basis works well for representing stationary signal segments, wavelet, chirplet, and Gabor bases can represent nonstationary phenomena such as modulation and transients more efficiently. The term physical model is here used to emphasise that the discretisation of continuous parameters is often in direct contradiction with the physics behind the generation of most real-world signals [60]. For example, frequencies and direction of arrivals are usually continuous quantities in nature. Consequently, the models for sparse decompositions are not necessarily as sparse as models based on the physical model as exemplified in [88]. As we have demonstrated in Paper F and Paper G, this effect becomes even more pronounced when the popular sampling strategy referred to as compressive sensing [34–36, 84] is used.

2.3 Models for the Noise Term

The noise term is usually modelled as either white Gaussian noise (WGN) or coloured (or non-white) Gaussian noise. WGN is by far the most popular model for the noise term, and most researchers and scientists attribute its ubiquitous use to the central limit theorem, but also properties such as mathematical tractability, geometrical invariance, and entropy maximisation favour the WGN model [134]. However, the WGN model is also frequently criticised for being too simple and unrealistic [210]. This criticism is usually based on a physical interpretation of the noise [124], and a noise model for coloured noise such as an ARMA model is therefore used instead to model the correlation in the noise. As argued by Jaynes and Bretthorst in [29, 30, 124], [125, ch. 7] and exemplified in [164], however, the simple WGN model should be used instead of a more complex noise model modelling correlations if it is not known whether the noise contains a correlation structure or not. The argument is based on the maximum entropy principle which states that the white Gaussian distribution has the largest entropy among all distributions when the location and the scale of the noise are constrained to some finite values [30, 134]. A related argument is that the white Gaussian distribution

minimises the Fisher information [134] and, consequently, maximises the Cramér-Rao lower bound [210]. Including constraints on the correlation structure lowers the entropy of the noise distribution so the WGN model is therefore a more general noise model in the sense that it is based on the least informative probability distribution [30].

An alternative to modelling coloured noise is to pass the data through a pre-whitening filter and then use a WGN model in the analysis of the filtered data [51, pp. 84-85]. This approach is especially attractive in frequency estimation since linear filtering does not change the frequency of a sinusoid.

2.4 Sampling Schemes

So far, most of the discussion has been applicable to both continuous-time and discrete-time signals. In Sec. 3, however, we exclusively work with a finite data set $\{x(t_n)\}_{n=0}^{N-1}$ originating from either a discrete-time signal or a sampled continuous-time signal. For the latter, the data set consists of N samples acquired at the times $\{t_n\}_{n=0}^{N-1}$ which we without loss of generality assume are ordered in time. Usually, the sampling times are uniformly spaced by a sampling period of T_s, whose reciprocal is the sampling frequency f_s, and this sampling scheme is referred to as uniform or periodic sampling. According to the famous Nyquist–Shannon sampling theorem [174, 204], the continuous-time signal can be exactly reconstructed from its discrete-time representation if its bandwidth does not exceed the Nyquist frequency which is half the sampling frequency. For nonuniform sampling, exact reconstruction is also possible provided that the average sampling frequency is at least twice the bandwidth, the Nyquist rate, of the continuous-time signal [31, 144, 165, 234]. Uniformly sampled signals are often written in terms of the sampling index n instead of the sampling time t_n. Since $t_n = T_s n + t_0$, this change of notation yields

$$\omega_i t_n = \tilde{\omega}_i n + \omega_i t_0 \tag{17}$$

where $\tilde{\omega}_i \triangleq \omega_i / f_s$ is the so-called digital angular frequency measured in radians per sample. The time t_0 is usually selected arbitrarily since its value translates into a change of the phase. Perhaps counter-intuitive, however, the value of $t_0 = -T_s(N-1)/2$ can be shown to yield the lowest estimation bound on any unbiased estimator of the phase of a sinusoidal signal [83].

The sampling strategy known as compressed sensing or compressive sampling [34–36, 84] has recently received a lot of attention since it allows perfect reconstruction of some signals sampled at a significantly lower rate than the Nyquist rate. For noiseless signals, which are S-sparse in a dictionary with D columns, perfect reconstruction is possible for a down-sampling factor up to the order of $(S/N)\log(D/S)$ [33]. For most sparse signals, $S \ll N$ and the sampling rate can therefore be reduced dramatically. For noisy signals or signals which are not exactly sparse in the dictionary, the reconstruction is only approximate, and in Paper F we have quantified the reconstruction error in terms of the down-sampling factor.

3 Inference

Statistical inference is the art of extracting information about the parameters $\boldsymbol{\theta}_k$ of a hypothesised model $\mathcal{M}_k$ and about the model itself from a finite data set $\boldsymbol{x}$. In this section, a brief overview is given over some of the fundamental inference methods for the sinusoidal models. Due to the large body of scientific literature published about this inference problem, the overview is limited to the basic model and the harmonic model. We have mainly focused on these two models since the inference methods for these two models are often so fundamental that they are used as a component in the inference methods for more complex models. Like most of the literature, the overview focuses on the zero-mean complex sinusoidal model sampled at a uniform sampling frequency, but most of the methods are applicable to both real- and complex-valued data.

3.1 Fourier-based Methods

Although astronomers in ancient times have tried to determine the length of the year or the period of the moon, the history of modern spectral analysis has its roots in Newton's work on optics which in the middle of the nineteenth century spurred significant interest in spectral analysis [133, 162, 190]. The first major advancement was made in the late nineteenth century by the introduction of the Schuster's periodogram [201] given by[1]

$$\hat{P}_X(\omega) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x(t_n)\exp(-j\omega t_n)\right|^2 \tag{18}$$

which was based on the Fourier transform introduced by Fourier in the early nineteenth century [92]. Until Yule in 1927 introduced his autoregressive method for spectrum analysis [235], the periodogram was the only numerical method for spectral analysis [190]. Despite its good resolution, the periodogram is an inconsistent estimator of the PSD since its variance does not decrease with an increasing number of data points N [215, Ch. 2]. To reduce this statistical variability, several modifications of the periodogram have been proposed such as Bartlett's method [9, 10], Welch's method [232], Daniell's method [72], the cepstrum [22], and most famously the Blackman-Tukey method [20]. These methods are different ways of averaging the periodogram or applying smooth lag-windows to the data. Due to the bias-variance trade-off of the non-parametric methods [131, p. 4], the cost of reducing the variance is a coarser resolution. Despite this trade-off, the Fourier-based methods have been and still are used extensively to estimate periodicities in data sets [203]. One major reason for this is the fast Fourier transforms (FFTs) [69] which have enabled very efficient implementations of the Fourier-based methods for uniformly sampled data. For non-uniformly sampled data, several

[1] Note that this definition of the periodogram gives a PSD estimate measured in watts per cycles per sample. If the periodogram estimate is divided by the sampling frequency f_s, the unit will be W/Hz [133].

approximate algorithms exist which require the same order of floating point operations as the FFTs [141, 180].

3.2 Methods from Classical Estimation Theory

Parallel to the development of the Fourier-based methods, frequency estimation methods based on probability theory were also developed [29]. The first methods were based on the least squares (LS) principle first used by Legendre [149] and Gauss [96] to estimate the orbit of astronomical objects. The LS principle is a method for estimating the model parameters of a postulated model in noise, and in frequency estimation the model is the sum of a number of sinusoids. For the complex basic model in Eq. (9), the model is in matrix notation

$$\boldsymbol{x} = \boldsymbol{Z}_k \boldsymbol{\alpha}_k + \boldsymbol{e} \tag{19}$$

where $\boldsymbol{\alpha}_k$ contains the complex amplitudes and the i'th column of $\boldsymbol{Z}_k$ is given by $\boldsymbol{z}_{i,k} = \begin{bmatrix} \exp(j\omega_i t_0) & \cdots & \exp(j\omega_i t_{N-1}) \end{bmatrix}^T$. Since the frequencies of these sinusoids are nonlinear parameters, the estimation of these frequencies using the LS principle is typically referred to as the nonlinear LS (NLS) method, and NLS estimates is given by [215, pp. 157–162]

$$\hat{\boldsymbol{\omega}}_k = \underset{\boldsymbol{\omega}_k \in \Omega_k}{\arg\max}\, \boldsymbol{x}^H \boldsymbol{Z}_k (\boldsymbol{Z}_k^H \boldsymbol{Z}_k)^{-1} \boldsymbol{Z}_k^H \boldsymbol{x} \tag{20}$$

where $(\cdot)^H$ denotes conjugate transposition and $\boldsymbol{\omega}_k$ contains the l_k frequencies with support Ω_k. To ensure identifiability, the support can be structured such that $0 \leq \omega_1 < \omega_2 < \ldots < \omega_l < 2\pi$. If just a single sinusoid is in the model, the NLS cost function in Eq. (20) is equivalent to the periodogram, and one of the first to notice this connection was Brunt in [32]. If the noise is white and has a Gaussian distribution, the NLS estimate is the value that maximises the sampling distribution or observation model $p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)$ where $\boldsymbol{\theta}_k$ and $\mathcal{M}_k$ denote a vector of unknown model parameters and the postulated $\mathcal{M}_k$, respectively. When the sampling distribution is viewed as a function of the unknown model parameters, the sampling distribution is no longer a probability distribution and is referred to as the likelihood function[2]. The maximiser of this function is therefore the maximum likelihood (ML) estimate. The link between the periodogram, the NLS principle, and ML enables an interesting interpretation of the periodogram. If the signal consists of a single sinusoid in WGN, the maximiser of the periodogram is the optimal estimator of the frequency in the absence of prior information. If this model is not true, however, the periodogram may give misleading answers [29, p. 20].

For multiple sinusoids, the NLS estimator is asymptotically the optimal unbiased estimator even for coloured noise [215, pp. 157–162]. Despite these good statistical properties, the NLS estimates of the frequencies are difficult to find in practice since

[2] Note that the two terms sampling distribution and likelihood function are often used synonymously.

the search for them involves a multi-dimensional maximisation of a multi-modal cost-function which is very sharply peaked around its global maximum [216]. The complexity of the estimation problem can be significantly reduced by breaking the multi-dimensional search into a cascade of lower-dimensional searches [59]. This can be done in various ways by for example matching pursuit [105, 108, 160], the RELAX algorithm [152, 211], and the expectation-maximisation principle [61, 79]. The disadvantage of using these simpler methods to find the frequency estimates is that they are suboptimal and therefore have worse statistical properties than the NLS estimates. However, their use can be justified by the fact that sinusoids are asymptotically orthogonal for any set of distinct frequencies [51, p. 30]. That is, $\lim_{N\to\infty} N(\boldsymbol{Z}_k^H \boldsymbol{Z}_k)^{-1} = \boldsymbol{I}_{l_k}$ where $\boldsymbol{I}_{l_k}$ is the $l_k \times l_k$ identity matrix[3]. If the l_k frequencies lie on the Fourier grid, the equality holds for any N, but for an arbitrary set of frequencies the relationship is only approximate for a finite N. Under this approximation, the NLS estimates in Eq. (20) can be written as

$$\hat{\boldsymbol{\omega}}_k \approx \underset{\boldsymbol{\omega}_k \in \Omega_k}{\arg\max}\, \boldsymbol{x}^H \boldsymbol{Z}_k \boldsymbol{Z}_k^H \boldsymbol{x} = \underset{\boldsymbol{\omega}_k \in \Omega_k}{\arg\max}\, \|\boldsymbol{Z}_k^H \boldsymbol{x}\|_2^2 = \underset{\boldsymbol{\omega}_k \in \Omega_k}{\arg\max} \sum_{i=1}^{l} \hat{P}_X(\omega_i) \ , \tag{21}$$

and these estimates are sometimes referred to as the approximate NLS (ANLS) estimates [51, Ch. 2]. From Eq. (21), we see that the ANLS estimates are simply the l_k largest peaks of the periodogram. Thus, provided that N is large and the frequencies are not too close to each other, 0 and 2π, the periodogram can be used to find good estimates of the frequencies. The ANLS estimates can also be used as initial values in a numerical optimisation algorithm which finds the NLS estimates in Eq. (20) using a line search or gradient methods [6, 26].

Optimal Estimators

Estimation theory is an area of statistics which is concerned with finding estimators of unknown parameters from noisy observations which contain some information about these parameters. An estimator is a function which maps the data $\boldsymbol{x}$ into an estimate $\hat{\boldsymbol{\theta}}_k = \boldsymbol{g}(\boldsymbol{x})$ of the unknown parameters $\boldsymbol{\theta}_k$. The function can be selected in an infinite number of ways and is usually selected as a trade-off between the statistical properties and the computational complexity with some examples being the minimum variance unbiased (MVU) estimators, best linear unbiased estimators (BLUE), ML estimators, LS estimators, and method of moments estimators [132]. The statistical performance of an estimator is usually quantified in terms of the risk function or the expected loss function [125, p. 410]

$$R(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k) = E[L(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k)] = \int_{\mathbb{C}^N} L(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k) p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k) d\boldsymbol{x} \tag{22}$$

[3] A similar limit holds in the real case.

where $E[\cdot]$ is the statistical expectation operator and $L(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k)$ is the loss incurred by guessing at $\hat{\boldsymbol{\theta}}_k$ when $\boldsymbol{\theta}_k$ is the true value. Minimising the risk function with respect to the estimator yields the useless result that the optimal unconstrained estimator is $\hat{\boldsymbol{\theta}}_k = \boldsymbol{\theta}_k$. Therefore, a practical optimal estimator must be constrained to be independent of the true parameter vector $\boldsymbol{\theta}_k$, but such an estimator may be hard to find or does not even exist for all values of $\boldsymbol{\theta}_k$ [132, Ch. 2]. The most popular way of measuring the loss is by the squared error measure. This loss function results in the mean squared error (MSE) risk function which can be decomposed as

$$R(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k) = E[\|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k\|_2^2] = \operatorname{tr}\left(\boldsymbol{C}_{\hat{\boldsymbol{\theta}}_k}\right) + \|\operatorname{bias}(\hat{\boldsymbol{\theta}}_k)\|^2 = \sum_{i=1}^{l_k}\left[\operatorname{Var}(\hat{\theta}_i) + \operatorname{bias}^2(\hat{\theta}_i)\right] \quad (23)$$

where the bias and variance of the estimator $\hat{\theta}_i$ are given by

$$\operatorname{bias}(\hat{\theta}_i) = E[\hat{\theta}_i] - \theta_i \quad (24)$$

$$\operatorname{Var}(\hat{\theta}_i) = E\left[(\hat{\theta}_i - E[\hat{\theta}_i])^2\right] = E[\hat{\theta}_i^2] - E^2[\hat{\theta}_i]\ . \quad (25)$$

Since only the variance is independent of $\boldsymbol{\theta}_k$, the search for optimal estimators is usually constrained to estimators having zero bias, the so-called unbiased estimators [132], and such unbiased estimators are MVU estimators if they minimise the MSE for all values of $\boldsymbol{\theta}_k$. Even if an MVU estimator exist, there exists no universal way to find it, but several approaches have been suggested which might produce it [132, Ch. 2–6].

Theoretical optimal estimators dependence on the true parameter vector is a consequence of the interpretation of probability in frequentist statistics. By interpreting probability as the relative frequency of occurrence after repeating an experiment an infinite number of times, the unknown parameters are deterministic variables [23, p. 5]. Consequently, the risk function's dependence on it cannot simply be removed by integrating over it. Another consequence of treating the unknown parameters as fixed quantities it that statistical statements cannot be made about the estimate, but only about the estimator which is judged by considered the long run performance over an infinite number of hypothetical repetitions of the experiment [23, pp. 5–6]. Although this difference might seem subtle, it is important to keep in mind when interpreting confidence intervals [193, p. 245], [172, p. 6] or performing hypothesis tests [157, Ch. 37]. In Sec. 3.5, the Bayesian approach to estimation is reviewed, and it does not suffer from these problems.

Cramér-Rao Lower Bound

The Cramér-Rao lower bound (CRLB) [71, 184] is a lower bound on the variance of any unbiased estimator, and it is frequently used as a benchmark tool. Provided that the support of the sampling distribution does not depend on $\boldsymbol{\theta}_k$ and that the sampling

distribution is differentiable [132, p. 67], the CRLB of an estimator is the diagonal elements of the Fisher information matrix (FIM) which is defined as [132, p. 529]

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_k) \triangleq E\left[\frac{\partial \ln p(\boldsymbol{x}|\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k^*}\left(\frac{\partial \ln p(\boldsymbol{x}|\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k^*}\right)^H\right] . \tag{26}$$

An unbiased estimator is efficient if its variance attains the CRLB for all values of $\boldsymbol{\theta}$. An example of an efficient estimator is the sample mean estimator of the mean of normally distributed data. The sample mean estimator is also the MVU estimator, but an MVU estimator is not necessarily efficient. For example, the sample variance estimator is the MVU estimator of the variance of normally distributed data, but not efficient. The ML estimator is asymptotically efficient if the CRLB exists and is finite. That is, for a large enough N, the ML estimator satisfies [132, p. 167]

$$\hat{\boldsymbol{\theta}}_k \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_k; \boldsymbol{\theta}_k, \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}_k)) \quad \text{as} \quad N \to \infty \tag{27}$$

where $\mathcal{N}(\cdot)$ denotes the normal distribution. This result and the fact that ML estimators can be found in a universal way explain the popularity of the ML estimator. However, since the ML estimate is not always easy to compute, it might be better to use simpler but suboptimal estimators in practical applications.

Unfortunately, the exact CRLB for the frequencies of l_k sinusoids cannot be found analytically, but only numerically. For a large N, uniformly sampled data, and WGN with variance σ^2, however, the CRLB of the frequencies of the complex basic model is approximately [131, Sec. 13.4]

$$\operatorname{var}(\hat{\omega}_i) \approx \frac{6\sigma^2 f_s^2}{N^3 A_i^2} , \tag{28}$$

and the CRLB of the fundamental frequency of the complex harmonic model of a perfectly periodic signal is approximately [55, 171]

$$\operatorname{var}(\hat{\omega}_0) \approx \frac{6\sigma^2 f_s^2}{N^3 \sum_{i=1}^{l_k} A_i^2 i^2} . \tag{29}$$

3.3 Subspace-based Methods

Subspace-based methods are an attractive alternative to the ML-based estimation methods since they can attain nearly the same estimation performance for time-series[4] as the NLS estimator without being based on the intractable cost function in Eq. (20). Consequently, the subspace-based methods might be significantly faster than the ML-based methods. The subspace-based methods were primarily developed as a solution

[4] For multi-channel signals, the MUSIC method is an asymptotically efficient estimator of the direction of arrival [222, Sec. 9.3.2].

to the direction of arrival (DOA) estimation problem in array signal processing [139], but they have also been applied extensively to frequency estimation in univariate time-series [51, 131, 215]. The first subspace-based method was Pisarenko's harmonic decomposition [178], which was later generalised in the multiple signal classification (MUSIC) method [18, 200]. Other popular subspace-based methods include the estimation of signal parameters via rotational invariance techniques (ESPRIT) [177, 195], the min-norm method [140], and weighted subspace fitting (WSF) [229]. The term *subspace-based* refers to the partitioning of the observed data into a signal subspace and a noise subspace. For the complex basic model in Eq. (9) with WGN and uniformly sampled zero-mean data partitioned into m-dimensional data vectors $\{\boldsymbol{x}(n)\}_{n=m-1}^{N-1}$ with $m < N$[5], the covariance matrix of $\boldsymbol{x}(n)$ is given by

$$\boldsymbol{C}_X = E[\boldsymbol{x}(n)\boldsymbol{x}^H(n)] = \boldsymbol{A}_k \boldsymbol{P}_k \boldsymbol{A}_k^H + \sigma^2 \boldsymbol{I}_m = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^H \tag{30}$$

where $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^H$ is the eigenvalue decomposition of $\boldsymbol{C}_X$, $\boldsymbol{P}_k = E[\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^H]$, and the i'th column of $\boldsymbol{A}_k$ is given by $\boldsymbol{a}(\omega_i) = \begin{bmatrix} 1 & \exp(j\omega_i) & \cdots & \exp(j\omega_i(m-1)) \end{bmatrix}^T$. When the phases of the sinusoids are assumed to be uncorrelated and uniform random variables[6], $\boldsymbol{P}_k$ is a diagonal matrix and the decomposition in Eq. (30) is termed the covariance matrix model. The eigenvectors in $\boldsymbol{U}$ are without loss of generality assumed to be sorted according to the eigenvalues in descenting order. The orthogonal matrix $\boldsymbol{U}$ can then be partitioned into $\boldsymbol{U} = \begin{bmatrix} \boldsymbol{S}_k & \boldsymbol{G}_k \end{bmatrix}$ where $\boldsymbol{S}_k$ and $\boldsymbol{G}_k$ are the $M \times l_k$-dimensional signal subspace and the $m \times (m - l_k)$-dimensional noise subspace. Since the noise subspace is orthogonal to the signal subspace, which is also spanned by $\boldsymbol{A}_k$, the frequencies of the l_k sinusoids satisfy that $\boldsymbol{A}_k^H \boldsymbol{G}_k = \boldsymbol{0}$ [215, p. 167]. In practice, $\boldsymbol{G}_k$ is unknown and must be estimated from the data by computing the eigenvalue decomposition of a covariance matrix estimate. The spectral-MUSIC [18, 200] estimates of the frequencies are therefore the l_k largest peaks of the so-called MUSIC pseudo-spectrum

$$\tilde{P}_X(\omega) = \|\hat{\boldsymbol{G}}_k^H \boldsymbol{a}(\omega)\|_2^{-2} \, . \tag{31}$$

An example of the MUSIC pseudo-spectrum is shown in Fig. 5. Since the direct maximisation of the pseudo-spectrum is a non-convex problem, the MUSIC frequency estimates are sometimes found via the root-MUSIC method [8] which finds the MUSIC estimates via polynomial rooting methods. Pisarenko's harmonic decomposition is equivalent to MUSIC for $M = l_k + 1$, but it is much less accurate than MUSIC for $l_k \ll m < N$ [215, p. 169]. The standard MUSIC method may produce frequency estimates which are far away from the true frequencies. Such spurious frequency estimates are avoided in a variation of the standard MUSIC method called the modified MUSIC

[5] For multi-channel signals, m is the number of sensors and $\boldsymbol{x}(n)$ contains the data from these sensors sampled at time index n

[6] Note that this assumption is not consistent with the interpretation of probability in frequentist statistics.

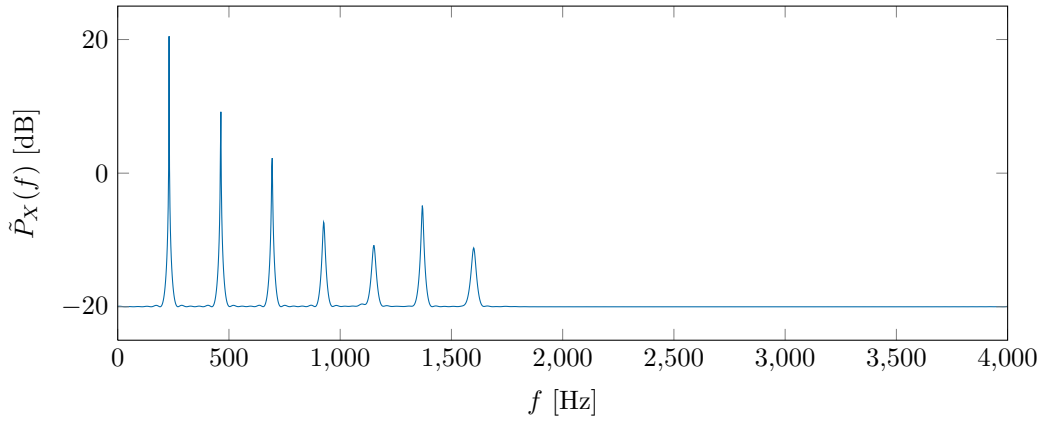

Fig. 5: The MUSIC pseudo-spectrum for the speech data in Fig. 2 when the signal subspace has seven eigenvectors.

method [219], but its statistical properties are slightly worse than that of the standard MUSIC method [215, p. 175].

The ESPRIT method is a very popular alternative to the MUSIC methods since its statistical properties are slightly better than that of the standard MUSIC method at a similar computational cost and, like the root-MUSIC method, it also finds the frequency estimates without solving a non-convex maximisation problem [215, p. 175]. By exploiting the shift-invariance property of the matrix $\boldsymbol{A}_k$, it can be shown that the angle of the eigenvalues of $(\boldsymbol{S}_{1,k}^H \boldsymbol{S}_{1,k})^{-1} \boldsymbol{S}_{1,k}^H \boldsymbol{S}_{2,k}$ are the frequencies which satisfy $\boldsymbol{A}_k^H \boldsymbol{G}_k = \boldsymbol{0}$. The matrices $\boldsymbol{S}_{1,k}$ and $\boldsymbol{S}_{2,k}$ are formed by removing the last and first row of $\boldsymbol{S}_k$, respectively, and in practice they must be computed from an eigenvalue decomposition of an estimated covariance matrix. Note that ESPRIT may produce spurious frequency estimates for real-world signals [54], [51, pp. 107–109].

Unlike the ML-based methods, the estimation accuracy of the subspace methods depends critically on the WGN assumption. For coloured noise, the data should therefore be pre-whitened prior to estimating the frequencies [51, pp. 84–85]. The eigenvalue decomposition of the estimated covariance matrix constitutes the major contribution to the computational cost of running the subspace-based methods. Consequently, several subspace tracking algorithms with a much lower computational complexity have been proposed and an overview over some of these are given in [68] and [87].

3.4 Filtering Methods

Like the subspace-based methods, the filtering methods have their roots in array signal processing where they are usually referred to as *beamforming* methods [139, 222, 226].

The basic idea in the filtering methods is to pass the signal through a filter or a filter bank which is designed to enhance or attenuate the sinusoids in the signal. Typically, the frequency estimates are then the set of frequencies which either maximise or minimise the power of the filtered output signal. The perhaps most popular filtering technique for frequency estimation is the Capon beamforming algorithm which is also known as the minimum variance distortionless response (MVDR) approach [37, 142]. For an FIR-filter with the m-dimensional impulse response vector $\boldsymbol{h}$, the power of the filtered output signal is

$$E[|\boldsymbol{h}^H \boldsymbol{x}(n)|^2] = \boldsymbol{h}^H \boldsymbol{C}_X \boldsymbol{h} \tag{32}$$

where $\boldsymbol{C}_X$ is the covariance matrix of $\boldsymbol{x}(n)$. Minimising this power subject to the constraint that the filter passes the frequency content at the frequencies $\boldsymbol{\omega}_k$ unaltered leads to the following equality constrained optimisation problem for the filter coefficients[7]

$$\hat{\boldsymbol{h}} = \underset{\boldsymbol{h} \in \mathbb{C}^m}{\arg\min}\, \boldsymbol{h}^H \boldsymbol{C}_X \boldsymbol{h} \quad \text{s. t.} \quad \boldsymbol{h}^H \boldsymbol{A}_k = \boldsymbol{1}_k^T \tag{33}$$

where $\boldsymbol{1}_k$ is an l_k-dimensional vector of ones. The optimisation problem is readily solved using the method of Lagrange multipliers [6, 26] and leads to the frequency estimates [57]

$$\hat{\boldsymbol{\omega}}_k = \underset{\boldsymbol{\omega}_k \in \Omega_k}{\arg\max}\, \boldsymbol{1}_k^T \left(\boldsymbol{A}_k^H \boldsymbol{C}_X^{-1} \boldsymbol{A}_k \right)^{-1} \boldsymbol{1}_k \ . \tag{34}$$

Although the filtering methods generally do not result in efficient estimators, they have good statistical properties and work well under adverse conditions such as closely spaced sinusoids [61]. The computational complexity of the estimator in Eq. (34) may be too high, primarily due to the multi-dimensional search over a non-convex cost function which may have very narrow peaks [51, p. 80]. However, the computational complexity may be reduced considerably by replacing the matrix $\boldsymbol{A}_k$ with the vector $\boldsymbol{a}(\omega)$ in Eq. (34). In this case, the frequency estimates are the l_k largest peaks of a one-dimensional cost function. This method is the Capon method, and it works very well in practice since the filter is designed in a data-adaptive way and thus suppresses interfering sinusoids and coloured noise. An example of the amplitude response of the Capon filter is shown in Fig. 6, and by comparing it to Fig. 3, we see that the Capon filter places nulls at the positions of the interfering sinusoids. A very efficient implementations of the Capon filter have also been developed [100, 148]. The Capon methods is often classified as a non-parametric estimator [215, Ch. 5], and for the special case of $m = 1$ it reduces to the periodogram. In addition to frequency estimation, the Capon method is also often used for PSD estimation [215, pp. 240–241] and amplitude estimation [215, p. 258]. However, extensive analysis and simulation studies have shown that the Capon estimates of the PSD and the amplitudes are biased towards

[7] The optimisation problem can also be formulated in terms of an entire filterbank instead of just a single filter [57].

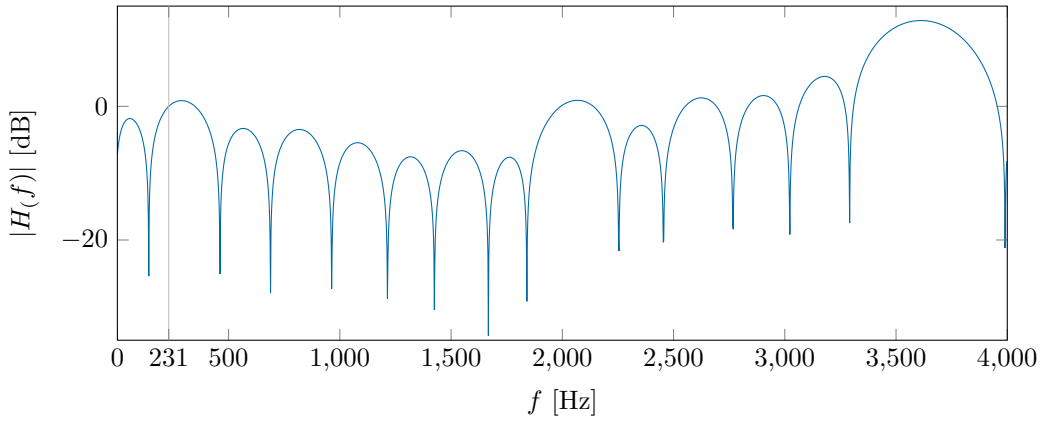

Fig. 6: The amplitude response of the Capon filter for the speech data in Fig. 2. The vertical line marks the frequency constraint. From Fig. 3, we see that the Capon filter places nulls in its amplitude response at the location of interfering sinusoids to minimise the power of the filter output.

zero [147, 150, 151, 212, 213]. The alternative filtering method called the amplitude and phase estimation (APES) method [151, 214] does not suffer from this bias and is therefore often used instead. Since the frequency resolution of APES is slightly worse than that of Capon, a combination of the two methods known as CAPES [120] can be used for the joint estimation of the frequencies and amplitudes. The APES method can also be used for the simultaneous estimation of multiple sinusoids analogue to the estimator in Eq. (34) [128].

The estimation performance of the filtering methods depends on the choice of the filter length m. For the Capon method, the filter length is chosen as a trade-off between resolution and bias, and in [213] a filter length in the interval $N/8 < m < N/4$ is recommended. Since the APES filter is unbiased, a maximum filter length of $m = \lfloor N/2 \rfloor$ is therefore recommended to maximise the resolution. However, we have observed that the APES amplitude spectrum for this maximum filter length suffers from ringing artefacts and overshoots. These phenomena can be significantly reduced by reducing the filter length or by using the so-called forward-backward estimate of the covariance matrix [121]. This covariance matrix estimate also significantly reduces the bias of the Capon method [150, 151]. Ideally, the filter length should be chosen in a data-adaptive way so short filters are used in the case of simple signals, which can be approximately described as a single sinusoid at a known candidate frequency in WGN, and long filters for complex signals. In Paper E, we have proposed a very simple way of doing this, and the resulting estimator is demonstrated to reduce the bias of the Capon method significantly with only a slight decrease in the resolution.

3.5 Bayesian Methods

None of the methods listed above give a direct indication of the accuracy of the frequency estimate for the observed data set. Monte Carlo simulations [189] and bounds such as the CRLB can be used to assess and benchmark the accuracy of an estimator across an infinite number of similar data sets, but similar statements cannot be made about the estimate itself. As discussed previously in Sec. 3.2, this is a consequence of interpreting a probability as a long run relative frequency of occurrences. On the other hand, the accuracy of estimates can be directly assessed in Bayesian statistics which is based on an alternative interpretation of probabilities. This interpretation is the oldest interpretation and was developed by Bernoulli, Bayes, and in particular by Laplace in the eighteens and early nineteenth centuries [206, Sec. 1.4]. In his famous treatise from 1812 [145], Laplace developed many of the fundamental results of modern probability theory [123]. Bayes and Laplace advocated interpreting a probability as an abstract quantity representing a state of knowledge or a degree of belief. As a consequence of this interpretation, unknown parameters were treated as random variables and not fixed quantities. This is reflected in one of Laplace's most important results, the Bayes' theorem, which is given by

$$p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{p(\boldsymbol{x}|\mathcal{M}_k)} \tag{35}$$

where $p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k)$, $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$, and $p(\boldsymbol{x}|\mathcal{M}_k)$ are usually referred to as the posterior distribution, the prior distribution, and the evidence, respectively. However, the Bayesian interpretation was initially rejected by many scientist as being too subjective and vague, and they instead promoted the physical interpretation of a probability as a long run relative frequency [206, Sec. 1.4]. First when Jeffreys [126] and Cox [70] rediscovered and promoted Bayes' and Laplace's rationale in the middle of the twentieth century, the Bayesian interpretation of probabilities again gained some interest and spurred some intense philosophical debates among statisticians [123].

In 1987, Jaynes [124] linked the periodogram with probability theory in a Bayesian framework. In the absence of prior information, he showed that the posterior distribution on the frequency of a single sinusoid in WGN was linked in a simple way to the periodogram. For complex data, this relationship is given by

$$p(\omega|\boldsymbol{x}, \mathcal{M}_k) \propto \left[1 - \frac{\hat{P}_X(\omega)}{\boldsymbol{x}^H\boldsymbol{x}}\right]^{-(N-1)} \tag{36}$$

where $\propto$ denotes proportional to. The posterior distribution is the complete answer to the inference problem about the frequency, and it can be used to find point estimates, compute probability intervals around these estimates, and test hypothesis. For example, as we have shown in Paper B, the optimal frequency estimate is the peak of the

periodogram and the variance of this frequency estimate is approximately

$$\text{var}(\hat{\omega}) \approx \frac{6\hat{\sigma}^2 f_s^2}{N^3 \hat{A}_i^2} \tag{37}$$

where $\hat{\sigma}^2$ and $\hat{A}_i$ are the ML estimates of the noise variance and amplitude, respectively. Note that the approximate variance of the frequency estimate is identical to the approximate CRLB in Eq. (28) with the true values for the noise variance and the amplitude replaced by their estimates. In Fig. 7, the posterior density of the frequency is shown. It consists of a single important peak whose shape is very similar to that of a Gaussian density. The work by Jaynes was extended significantly by Bretthorst in [29] where he considered multiple sinusoids and model comparison. Since the main purpose of the work by Jaynes and Bretthorst was to give insight into spectral estimation in a Bayesian framework and to facilitate an easy interpretation of the results, they derived their results using analytical approximations under the assumption that the sinusoids were well-separated and enough data were available. In for example [4, 85, 86, 196] numerical techniques were used instead of analytical approximations to get more accurate results and to avoid making these simplifying assumptions. Unfortunately, the use of numerical techniques often increases the computational complexity of the algorithms significantly. Besides the basic sinusoidal model in Eq. (8) and Eq. (9), Bayesian inference has been applied to numerous other sinusoidal models in various contexts. For example the harmonic models have received some attention in [73, 75, 101, 102] and models containing amplitude and/or frequency modulation have received some attention in [39, 40, 42, 113, 136, 223, 224]. In Paper B, we have extended the work by Jaynes and Bretthorst to real- and complex-valued harmonic signals.

Optimal Bayesian Point Estimates

Through Bayes' theorem in Eq. (35), the posterior distribution on a parameter $\boldsymbol{\theta}_k$ optimally combines the prior information about the parameter with the information learned about the parameter by observing the data $\boldsymbol{x}$ [103, p. 75]. By optimal, we mean that the posterior distribution contains all the information about the unknown parameter. Moreover, for some loss function, a corresponding point on the posterior distribution minimises the Bayes' risk function and is therefore an optimal estimate. The Bayes' risk function is directly linked to the risk function previously considered in Sec. 3.2. Since the unknown parameter is a random variable in a Bayesian framework, it can be integrated out of the risk function to obtain the Bayes' risk function

$$R(\hat{\boldsymbol{\theta}}_k) = \int_{\Theta_k} R(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k) p(\boldsymbol{\theta}_k | \mathcal{M}_k) d\boldsymbol{\theta}_k \tag{38}$$

$$= \int_{\mathbb{C}^N} \left[\int_{\Theta_k} L(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k) p(\boldsymbol{\theta}_k | \boldsymbol{x}, \mathcal{M}_k) d\boldsymbol{\theta}_k \right] p(\boldsymbol{x} | \mathcal{M}_k) d\boldsymbol{x} \tag{39}$$

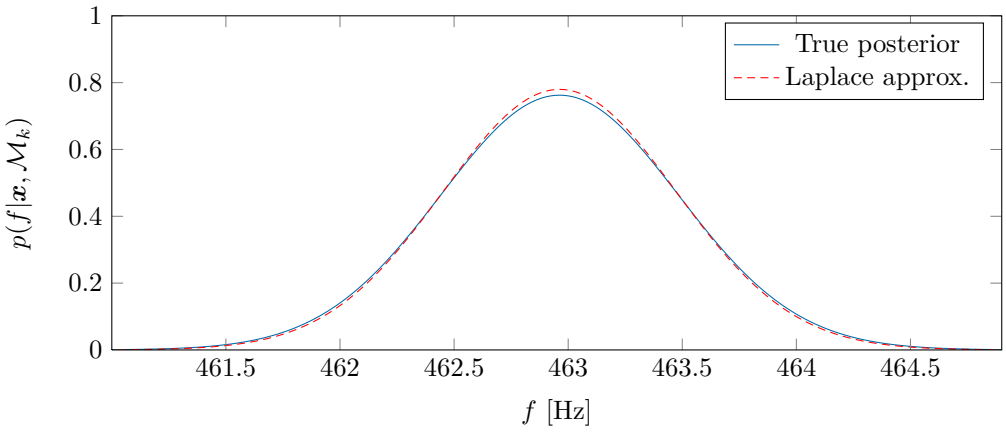

Fig. 7: The posterior distribution with the density given by Eq. (36) and its Laplace approximation for the speech data in Fig. 2. The standard deviation is approximately 0.5 Hz.

where Θ_k is the support of $\boldsymbol{\theta}_k$. In contrast to $R(\boldsymbol{\theta}_k, \hat{\boldsymbol{\theta}}_k)$, the Bayes' risk function $R(\hat{\boldsymbol{\theta}}_k)$ does not depend on the true parameter. Moreover, since $p(\boldsymbol{x}|\mathcal{M}_k) \geq 0$, the Bayes' risk can be minimised by minimising the inner integral. For example, the mean, the mode and the median of the posterior distribution minimises the Bayes' risk for the squared error loss function, the hit-or-miss loss function, and the absolute error loss function, respectively [132, Sec. 11.3]. In a Bayesian framework, a universal method to find the optimal estimator therefore exists, and it does not depend on the true value of the parameter. However, despite that this optimal estimator is conceptually easy to find, it might be hard or impossible in practice to find an analytical expression for the posterior distribution and its moments, mode and median.

Practical Problems

Despite the conceptual simplicity of Bayesian inference, there are two main practical problems associated with it. The first problem is the elicitation of the prior distribution. How does one turn prior information into a probability distribution in a consistent way, and is it possible to assign prior distributions in an objective fashion in the absence of prior information? There are two popular answers to these questions with roots in the objective Bayesian methods [12, 122, 125] and the subjective Bayesian methods [16, 77], respectively. In the objective view, two different researchers (and even a robot) should assign the same prior distribution if they had the same information, and the prior distributions are usually elicited using maximum entropy methods [122], invariance of transformation groups [126], and reference analysis [16, Sec. 5.4], [13]. The resulting prior distributions are often referred to as uninformative priors, and they are

often improper as they do not integrate to one. A famous example of such a prior is the Jeffreys' prior [126]. Improper priors might pose significant problems for the purpose of performing model comparison [14]. In the subjective view, prior distributions are assigned by experts or according to one's personal belief. However, in inference problems involving a lot of parameters and possibly several candidate models, it might be infeasible or impractical to do the assignment in this way.

The second practical problem is the evaluation of high-dimensional and intractable integrals which must be evaluated to remove nuisance parameters[8], to find moments such as the mean, and to find the evidence. Such an integral can be written as

$$E[f(\boldsymbol{y})] = \int_Y f(\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y}, \tag{40}$$

and approximately evaluated using Monte Carlo integration in which T samples $\{\boldsymbol{y}^{[\tau]}\}_{\tau=1}^{T}$ are sampled from the distribution with density $p(\boldsymbol{y})$ and inserted into the sum

$$E[f(\boldsymbol{y})] \approx \frac{1}{T}\sum_{\tau=1}^{T} f(\boldsymbol{y}^{[\tau]}) \ . \tag{41}$$

This sum converges with an increasing T to the integral in Eq. (40) if $p(\boldsymbol{y})$ has a finite variance. Note that the convergence speed does not depend on the size of $\boldsymbol{y}$, but on the correlation between the generated samples from $p(\boldsymbol{y})$ [19, p. 524]. Often, however, independent samples cannot easily be generated from the desired density $p(\boldsymbol{y})$. In the 1980s, the Bayesian community therefore adopted the stochastic integration methods known as Markov chain Monte Carlo (MCMC) methods which have their origin in physics [19, p. 538]. Along with the rapid increase in computational power, the MCMC methods revolutionised Bayesian statistics and made inference in high-dimensional problems feasible. The MCMC methods work by selecting the transition kernel of an ergodic Markov chain[9] such that the invariant distribution of the Markov chain is the desired distribution which we wish to draw samples from. After an initial transient period in which the Markov chain converges, samples generated by the Markov chain are distributed according to the desired distribution. The two most well-known MCMC methods are the Metropolis-Hastings (MH) algorithm [46, 114, 169] and the Gibbs sampler [98], but many other methods exist [19, 97, 99, 157, 189].

Despite the rapid development in the MCMC methods, the resulting algorithms often suffer from high computational complexity. A more efficient alternative to these stochastic methods are the deterministic methods which are based on analytical approximations. The deterministic methods are sometimes called variational inference techniques, and they work by assuming a particular parametrisation or factorisation of

[8] A nuisance parameter is an unknown parameter which we are not interested in.

[9] An ergodic Markov chain converges to the required invariant distribution for any initial configuration [19, p. 540].

the desired distribution [19, Ch. 10]. The most popular analytical approximation is the Laplace approximation [146, 221] which can be used to approximate both single- and multi-modal distributions with a single or a mixture of normal distributions [97, Ch. 12]. In Fig. 7, the density of the Laplace approximation to the posterior distribution on the frequency is shown. Clearly, the true density has slightly heavier tails than its Laplace approximation, but for most practical applications the difference between the two densities is negligible.

3.6 Model Order Selection and Comparison

So far, the model $\mathcal{M}_k$ has been assumed to be known so that the inference problem was to estimate the unknown parameters $\boldsymbol{\theta}_k$ in this model. However, it is important to keep in mind that a model is typically just an approximate description of a much more complicated physical process which we might not even understand to its full extend. In 1987, Box stated this as, *"Essentially, all models are wrong, but some are useful"* [24, p. 424]. Since the exact description of the physical process generating the data $\boldsymbol{x}$ is typically unknown, several candidate models might be proposed. An important inference problem is therefore to compare these models in the light of the observed data. Even if we are fairly confident that our data set consist of a sum of sinusoids in WGN, we often do not know how many sinusoids there are. For example, it is not obvious from Fig. 2 and Fig. 3 how many sinusoids should be included in a model for the data. It might be tempting to use a single model with a large number of sinusoids as this model contains the simpler models as special cases, but this is clearly not an efficient way of compressing a data set, and it might also lead to wrong estimates as we have demonstrated in Example 3.1.

Example 3.1 (Wrong Number of Harmonics)
Suppose a noisy periodic signal consisting of just the fundamental harmonic at the frequency ω_0 is observed. Moreover, assume that the noise is white and Gaussian with variance σ^2, that the number of harmonics is wrongly set to $l_k = 2$, and that the number of observations N is large enough to justify the approximation $N(\boldsymbol{Z}_k^H \boldsymbol{Z}_k)^{-1} \approx \boldsymbol{I}_{l_k}$. In Paper B, we have shown that the estimate of the fundamental frequency maximises the cost function

$$C_2(\omega) \approx N^{-1} \boldsymbol{x}^H \boldsymbol{Z}_2(\omega) \boldsymbol{Z}_2^H(\omega) \boldsymbol{x} = N^{-1} \sum_{i=1}^{2} |\boldsymbol{z}_i^H(\omega)\boldsymbol{x}|^2 = N^{-1} \sum_{i=1}^{2} |\boldsymbol{z}_1^H(i\omega)\boldsymbol{x}|^2 \ . \quad (42)$$

It then follows from the covariance matrix model in Eq. (30) that the expected value of the cost function at ω_0 and $\omega_0/2$ is

$$E[C_2(\omega_0)] \approx N|\alpha_1|^2 + \sigma^2 + \sigma^2 = N|\alpha_1|^2 + 2\sigma^2 \quad (43)$$

$$E[C_2(\omega_0/2)] \approx \sigma^2 + N|\alpha_1|^2 + \sigma^2 = N|\alpha_1|^2 + 2\sigma^2 \ . \quad (44)$$

In all other frequency points, we have that $E[C_2(\omega)] \approx 2\sigma^2$. Thus, since $E[C_2(\omega_0)] \approx E[C_2(\omega_0/2)]$, we would get the so-called pitch halving problem with a probability of approximately 50 %. If we instead of $l_k = 2$ harmonics wrongly assumed l_k harmonics, we would have that

$$E[C_{l_k}(\omega_0/i)] \approx N|\alpha_1|^2 + l_k\sigma^2 \tag{45}$$

so that the estimate of the pitch is wrong with a probability of approximately $100(l_k - 1)/l_k$ %. On the other hand, if we could estimate the true model order, we would not suffer from problems with fractional estimates of the fundamental frequency.

For several decades, many model selection and comparison methods have been proposed, and a few good overviews over most of them might be found in [67, 185, 215, 217, 218]. The methods are based on various principles such as probability theory, cross-validation, prediction performance, coding theory, and principal component analysis. These methods can roughly be divided into three groups with the first group being those methods which require an a priori estimate of the model parameters, the second group being those methods which do not require such estimates, and the third group being those methods in which the model parameters and model are estimated and detected jointly [38]. In the rest of this section, we briefly review some of the methods in these three groups, and in Fig. 8, we have compared three of the methods.

Information Criteria

The large number of information criteria is perhaps the most prominent type of model selection method belonging to the first group of methods. They are used to find the most probable model index by solving an optimisation problem of the form [217]

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \left[2\ln p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k) - \nu_k h(\nu_k, N)\right] \tag{46}$$

where $\hat{\boldsymbol{\theta}}_k$, $\mathcal{K}$, $h(\nu_k, N)$, and ν_k are the ML estimate of the model parameters, the set of model indices, the penalty coefficient, and the number of real-valued and free parameters in the model[10], respectively. The various information criteria differ in terms of how the penalty coefficient is selected. For example, for $h(\nu_k, N) = \{2, 2N/(N-\nu_k-1), \ln N\}$, we get the Akaike information criterion (AIC) [3], the corrected AIC (AIC_c) [119], and the original minimum description length (MDL) [186, 187], respectively, but many others exist [185]. The major advantage of the information criteria is that they are very simple to implement and therefore lead to fast algorithms. However, the simplicity of the criteria is often obtained by making approximations based on for example asymptotic properties, and this degrades the model selection performance and leads to problems

[10] The number of free parameters is not always the same as the number of unknown parameters. For an example, see [231].

with under- or overestimation of the model order and consistency issues [80, 199, 217, 231].

Principal Component Analysis

The second group of model selection and comparison methods are typically based on a principal component analysis of the covariance matrix of the data. An attractive property of these methods is that they do not rely on that the distribution of the data is known and that the ML estimate of the model parameters can easily be found, but only on that a consistent estimate of the covariance matrix is available [56]. Since the eigenvalues pertaining to the eigenvectors forming the noise subspace of the eigenvalue decomposition in Eq. (30) are all equal to the noise variance and smaller than the remaining eigenvalues, the model order can in principle be estimated by counting the number of eigenvalues equal to the noise variance. In practice, however, it is not easy to separate the eigenvalues of the estimated covariance matrix into signal and noise-subspace eigenvalues since the transition from one set to the other is often smooth. By comparing the ratio between the arithmetic and geometric means of the eigenvalues [231], [51, Sec. 4.5], an estimate of the model order can be found, but the estimator suffers from overestimation problems in the case of coloured noise [153, 233]. As an alternative to the eigenvectors, the eigenvalues can be used instead as in, e.g., the subspace-based automatic model order selection (SAMOS) method [176] and the estimation error (ESTER) method [7]. A related, but more general, idea is to measure how orthogonal the matrix $\boldsymbol{A}_k$ of Eq. (30) is to the eigenvectors in the noise subspace $\boldsymbol{G}_k$ for various choices of l_k. This methods is called the angle between subspaces (ABS) method [56, 104], and it selects the best model index by solving the following optimisation problem [56]

$$\hat{k} = \arg\min_{k \in \mathcal{K}} \sum_{k=1}^{l_k} \min_{\omega_k \in \Omega_k} \frac{\|\hat{\boldsymbol{G}}_k^H \boldsymbol{a}(\omega_k)\|_2^2}{\min(l_k, m - l_k)m} \tag{47}$$

where the nominator is the same as the MUSIC cost function and the inverse of the MUSIC pseudo-spectrum in Eq. (31).

Bayesian Methods

In the third group of methods, the Bayesian methods are found. These are conceptually very simple as they compute the posterior distribution over the set of candidate models via Bayes' theorem

$$p(\mathcal{M}_k|\boldsymbol{x}) \propto p(\boldsymbol{x}|\mathcal{M}_k)p(\mathcal{M}_k) \ . \tag{48}$$

The likelihood $p(\boldsymbol{x}|\mathcal{M}_k)$ is the evidence in Eq. (35) and given by the integral

$$p(\boldsymbol{x}|\mathcal{M}_k) = \int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k \ . \tag{49}$$

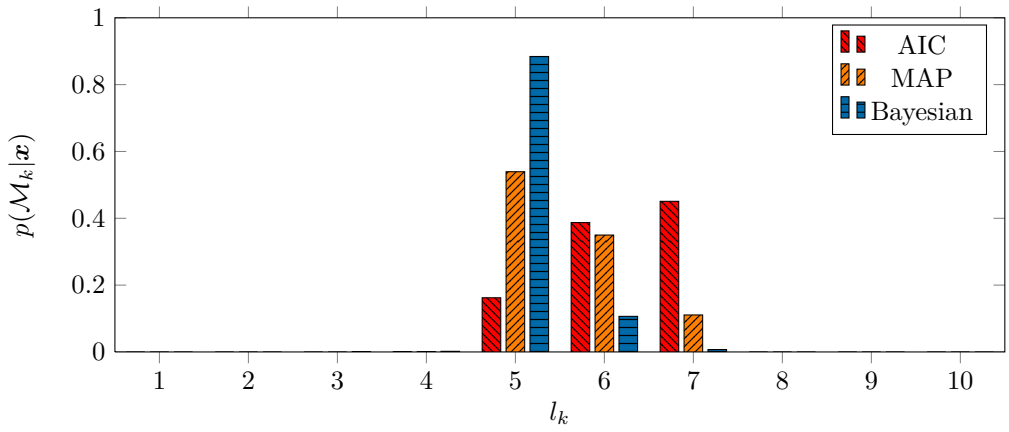

Fig. 8: Posterior distributions for the model order given the speech data in Fig. 2. The posterior distributions for the AIC and MAP methods are found using the approach described in [218], and the posterior distribution of the Bayesian method is the EB method from Paper B. The MUSIC methods with ABS [56] does not produce a posterior distribution, but selects a model order of $l_k = 2$.

Therefore, the evidence if often also referred to as the marginal likelihood. In addition the conceptual simplicity, the Bayesian methods also have numerous other advantages [14, 15]: Bayesian model comparison is consistent under very mild conditions, naturally selects the simplest model which explains the data reasonably well (the principle of Occam's razor [157, Ch. 28]), takes model uncertainty into account for estimation and prediction, works for non-nested models, and enables a more intuitive interpretation of the results. Unfortunately, these benifits are often dwarfed by the practical problems mentioned in Sec. 3.5 which are even more problematic in the context of model comparison. When several candidate models are considered, the MCMC methods must draw samples for the model index and the model parameters across all candidate models. The reversible jump MCMC method [107] is perhaps the most popular way of doing this, but it might be very hard to select an efficient transition kernel of the underlying Markov chain. Other popular stochastic integration techniques include importance sampling [5] and Chib's methods [44, 45], and in [109] and overview over some of the popular methods is given. The elicitation of prior distributions is also very important in the context of model comparison as the use of improper or uninformative prior distributions may lead to non-sensible answers. In Paper A, we have a more thorough discussion of these issues. To circumvent these problems, analytic approximations and asymptotic considerations can be used. The most popular example of this is the Bayesian information criterion [202] and the asymptotic maximum a posteriori (MAP) criteria [81, 82]. They are both based on a Laplace approximation to the marginal likelihood in Eq. (49), and they throw away first-order terms and use asymptotic consideration to avoid the

specification of proper prior distributions. The BIC method is equivalent to the MDL principle[11], and the MAP criteria may be seen as a generalisation of the BIC method to accommodate for the fact that every model parameter does not contribute with the same penalty to the overall model complexity penalty term. In Paper A, we extend the work by Djuric.

4 Contributions

As the title of this thesis suggests, the range of contributions is rather diverse. The individual contributions are therefore sometimes only connected by the framework described in this introductory part of the thesis. Paper A, Paper B, Paper E, and Paper G all consider model selection and comparison in both general and specific models. Paper A, Paper B, Paper C, and Paper D treat Bayesian inference in various parametric sinusoidal models. Whereas Paper C and Paper D are based on stochastic integration techniques, Paper B and Paper C are based on analytical approximations. The latter usually leads to less accurate results, but faster algorithms. Finally, Paper F and Paper G is concerned with quantifying the loss in estimation accuracy in the context of compressed sensing. Below is a short description of the contributions in each of the papers which constitute the main part of this thesis.

Paper A Performing model comparison and selection is in general very difficult, and, therefore, many simplifications are usually made. However, it is not always clear which assumptions these simplifications are based on, and how the simplifications are made. This is in particular the case in Djuric' asymptotic MAP criteria [82] in which it is not always obvious which terms can be safely ignored. In Paper A, we therefore extend the work by Djuric in several ways. First, we consider the general model comparison problem in a Bayesian framework and discuss the difficult elicitation of the prior distribution. These results are not new, but are scattered across mainly the statistical literature. Second, we derive a few new model selection criteria in a fairly general model, and we give some new insight into the implicit assumptions made in the various information criterion. Finally, we demonstrate that the proposed model comparison methods outperform the AIC, the BIC, and the MAP methods in terms of model selection accuracy and prediction performance.

Paper B Joint fundamental frequency estimation and model order detection is an important problem which has received a lot of research attention. However, most of the existing solutions are rather heuristic and very application specific. In Paper B, the problem is solved in a consistent fashion as outlined in Paper A

[11] We here refer to the original form of the MDL principle proposed by Rissanen in 1978 in [186]. Rissanen and others have since refined this principle in, e.g., [110, 188, 192].

when only a minimum of prior information is available. Moreover, the proposed solution is demonstrated to improve on what we believe is the state-of-the-art method [51, Sec. 2.6] for solving this problem. Although the method cannot be used as a stand alone pitch detection and tracking system in speech or music applications, we believe that it might be a useful component in such systems as well as in other application domains in which periodic signals must be analysed.

Paper C In this paper, we consider a dynamic sinusoidal model which is able to capture phenomena such as amplitude and frequency modulation. We show that the dynamic sinusoidal model models the evolution of the in-phase and quadrature components of the sinusoids by first-order Gauss-Markov processes. Using a Gibbs sampler, we derive an efficient algorithm which can be used to make inference about the unknown model parameters and to interpolate missing or corrupted observations. Finally, we use the algorithm for packet-loss concealment of both speech and music signals. The proposed method works well for interpolating short segments.

Paper D This paper is based on the results in Paper C, but also assess the quality of the packet-loss concealment via a MUSHRA listening test [1, 230] for various packet-loss probabilities and interpolation methods. These listening tests indicated that the proposed algorithm can be used to increase the fidelity of the degraded audio signals.

Paper E Although the Capon and APES filtering methods for amplitude estimation have been analysed extensively, it is still an open question if the filter length can be selected in some data-adaptive fashion. In Paper E, we have suggested a simple way of estimating the length of the Capon filter in a data-adaptive way. The method is based on Djuric' MAP criteria [82], and it is demonstrated to reduce the bias in the amplitude estimate significantly while still achieving nearly the same resolution.

Paper F Recently, compressed sensing has received a lot of research attention. The main reason for this is that it allows perfect reconstruction of under-sampled signals provided that they are sparse enough in some finite dictionary. However, when compressed sensing is used to sample for example sinusoidal signals, then perfect reconstruction is only possible if the true frequencies is exactly on the frequency grid defined by the dictionary. As the frequencies are typically continuous in nature, a loss in the reconstruction performance might be expected when signals are acquired using compressed sensing. In Paper F, we have quantified this loss for various sensing matrices in terms of the Cramér-Rao lower bound. The results show that the popular sensing matrices all lead to an expected loss in estimation accuracy proportional to the under-sampling factor.

Paper G This paper is an example of the points made in Paper F since compressed sensing is applied to the problem of joint direction-of-arrival and order estimation. The results are consistent with the observations in Paper F. They show that the estimation problem can be solved efficiently using compressed sensing, but that the estimation accuracy decreases with the down-sampling factor.

5 Conclusion

In this thesis, several new results within the framework of sinusoidal parameter estimation have been documented. The main contributions have been the analysis of the fundamental frequency and the model comparison problems in Paper B and Paper A which have resulted in a few new algorithms and facilitated a better understanding of the problems. We believe that these algorithms can be used as useful components in larger algorithms solving more complex problems in several application domains. Moreover, we have also demonstrated that these methods outperform similar state-of-the-art algorithms on synthetic signals while still having a tractable computational complexity.

Our analysis of the problems has primarily been performed in a Bayesian framework. Although some practical problems must be resolved when working in this framework, its conceptual simplicity, intuitive results, and ability to include prior information in a consistent way provide some strong arguments in favour of it. With the constant increase in computing power, the practical problem regarding the evaluation of intractable integrals will become less of an issue and allow us to use even more complex models.

Although the estimation of sinusoidal parameters from noisy observations has been subject to extensive research for several decades, we believe that more can be done in this area. For example, it would be interesting to use the Bayesian framework in connection with frame based processing of signals since this framework allows us to incorporate interframe dependencies in a consistent manner via the prior distribution. Another interesting extension for audio application would be to incorporate a perceptual distortion measure when doing, e.g., model comparison. For example, it is well-known that the human brain perceives a pitch at the same frequency as the fundamental frequency even though the first harmonic is missing.

References

[1] "Method for the subjective assessment of intermediate quality levels of coding systems," 2003, ITU BS.1534-1.

[2] J. Abesser, H. Lukashevich, and G. Schuller, "Feature-based extraction of plucking and expression styles of the electric bass guitar," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2010, pp. 2290–2293.

[3] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[4] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2667–2676, 1999.

[5] C. Andrieu, A. Doucet, and C. P. Robert, "Computational advances for and from Bayesian analysis," *Statist. Sci.*, vol. 19, no. 1, pp. 118–127, Feb. 2004.

[6] A. Antoniou and W.-S. Lu, *Practical Optimization: Algorithms and Engineering Applications*. Springer, Mar. 2007.

[7] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 450–458, Feb. 2006.

[8] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 8, Apr. 1983, pp. 336–339.

[9] M. S. Bartlett, "Smoothing periodograms from time-series with continuous spectra," *Nature*, vol. 161, pp. 686–687, May 1948.

[10] ——, "Periodogram analysis and continuous spectra," *Biometrika*, vol. 37, no. 1/2, pp. 1–16, Jun. 1950.

[11] H. Beigi, *Fundamentals of Speaker Recognition*. Springer, Dec. 2011.

[12] J. O. Berger, "The case for objective Bayesian analysis," *Bayesian Anal.*, vol. 1, no. 3, pp. 385–402, 2006.

[13] J. O. Berger, J. M. Bernardo, and D. Sun, "The formal definition of reference priors," *Ann. Stat.*, vol. 37, no. 2, pp. 905–938, Apr. 2009.

[14] J. O. Berger and L. R. Pericchi, "Objective Bayesian methods for model selection: Introduction and comparison," *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, vol. 38, pp. 135–207, 2001.

[15] ——, "The intrinsic Bayes factor for model selection and prediction," *J. Amer. Statistical Assoc.*, vol. 91, no. 433, pp. 109–122, Mar. 1996.

[16] J. M. Bernardo and A. Smith, *Bayesian Theory*, 1st ed. John Wiley and Sons Ltd, 1994.

[17] O. Besson, M. Ghogho, and A. Swami, "Parameter estimation for random amplitude chirp signals," *IEEE Trans. Signal Process.*, vol. 47, no. 12, pp. 3208–3219, Dec. 1999.

[18] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Apr. 1979, pp. 306–309.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, Aug. 2006.

[20] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra: From the Point of View of Communications Engineering*. Dover Publications, 1959.

[21] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 520–538, Apr. 1992.

[22] B. Bogert, M. Healy, and J. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proc. Symp. Time Series Analysis*, 1963, pp. 209–243.

[23] W. M. Bolstad, *Introduction to Bayesian Statistics*, 2nd ed. Wiley-Interscience, August 2007.

[24] G. E. P. Box and N. R. Draper, *Empirical model-building and response surface.* John Wiley & Sons, Inc., Jan. 1987.

[25] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis, Forecasting and Control*, 3rd ed. Printice-Hall International, Inc., 1994.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, Mar. 2004.

[27] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 110–120, Mar. 2004.

[28] ——, "Damped and delayed sinusoidal model for transient signals," *Signal Processing, IEEE Transactions on*, vol. 53, no. 5, pp. 1720–1730, May 2005.

[29] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation.* Springer-Verlag, Berlin Heidelberg, 1988.

[30] ——, "The near-irrelevance of sampling frequency distributions," in *Max. Entropy and Bayesian Methods*, 1999, pp. 21–46.

[31] ——, "Nonuniform sampling: Bandwidth and aliasing." *AIP Conf. Proc.*, vol. 567, no. 1, pp. 1–28, 2001.

[32] D. Brunt, *The Combination of Observations.* Cambridge University Press, 1931.

[33] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 1, pp. 59–73, Jul. 2011.

[34] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[35] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[36] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[37] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[38] T. Cassar, K. P. Camilleri, and S. G. Fabri, "Order estimation of multivariate ARMA models," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 3, pp. 494–503, Jun. 2010.

[39] A. T. Cemgil, "Bayesian Music Transcription," Ph.D. dissertation, Radboud University of Nijmegen, 2004.

[40] A. T. Cemgil and S. J. Godsill, "Efficient variational inference for the dynamic harmonic model," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2005, pp. 271–274.

[41] ——, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *Proc. European Signal Processing Conf.*, 2005.

[42] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[43] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, Jan. 2001.

[44] S. Chib, "Marginal likelihood from the Gibbs output," *J. Amer. Statistical Assoc.*, vol. 90, no. 432, pp. 1313–1321, Dec. 1995.

[45] S. Chib and I. Jeliazkov, "Marginal likelihood from the Metropolis-Hastings output," *J. Amer. Statistical Assoc.*, vol. 96, no. 453, pp. 270–281, Mar. 2001.

[46] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[47] C.-M. Cho and P. M. Djuric, "Bayesian detection and estimation of cisoids in colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2943–2952, Dec. 1995.

[48] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," 2012, unpublished manuscript.

[49] ——, "A method for low-delay pitch tracking and smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012.

[50] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Advances in Signal Process.*, vol. 13, Jun. 2011.

[51] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, B. H. Juang, Ed. Morgan & Claypool, 2009.

[52] M. G. Christensen, A. Jakobsson, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal signal decomposition for audio," *IEEE Signal Process. Lett.*, vol. 13, no. 7, pp. 389–392, Jul. 2006.

[53] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation using harmonic music," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2006, pp. 521–524.

[54] ——, "Fundamental frequency estimation using the shift-invariance property," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2007, pp. 631–635.

[55] ——, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1635–1644, Jul. 2007.

[56] ——, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. on Advances in Signal Process.*, 2009.

[57] M. G. Christensen, J. H. Jensen, A. Jakobsson, and S. H. Jensen, "On optimal filter designs for fundamental frequency estimation," *IEEE Signal Process. Lett.*, vol. 15, pp. 745–748, 2008.

[58] M. G. Christensen and S. H. Jensen, "Computationally efficient amplitude modulated sinusoidal audio coding using frequency-domain linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May 2006.

[59] ——, "On perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. Speech Audio Process.*, vol. 41, no. 1, pp. 99–109, Jan. 2006.

[60] M. G. Christensen and J. K. Nielsen, "Joint direction-of-arrival and order estimation in compressed sensing using angles between subspaces," in *Proc. IEEE Workshop on Stat. Signal Process.*, Jun. 2011, pp. 449–452.

[61] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[62] M. G. Christensen and S. van de Par, "Rate-distortion efficient amplitude modulated sinusoidal audio coding," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 2, Nov. 2004, pp. 2280–2284.

[63] ——, "Efficient parametric coding of transients," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1340–1351, Jul. 2006.

[64] M. G. Christensen, P. Vera-Candeas, S. D. Somasundaram, and A. Jakobsson, "Robust subspace-based fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2008, pp. 101–104.

[65] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal models for audio modeling and coding," in *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 2773. Springer-Verlag, Oct. 2003, pp. 1334–1342.

[66] M. Christensen and S. Jensen, "New results on perceptual distortion minimization and nonlinear least-squares frequency estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2239–2244, Sep. 2011.

[67] M. Clyde and E. I. George, "Model uncertainty," *Statist. Sci.*, vol. 19, no. 1, pp. 81–94, Feb. 2004.

[68] P. Comon and G. H. Golub, "Tracking a few extreme singular values and vectors in signal processing," *Proc. IEEE*, vol. 78, no. 8, pp. 1327–1343, Aug. 1990.

[69] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.*, vol. 19, no. 90, pp. 297–301, Apr. 1965. [Online]. Available: http://www.jstor.org/stable/2003354

[70] R. T. Cox, "Probability, frequency and reasonable expectation," *American Journal of Physics*, vol. 14, no. 1, pp. 1–13, Jan. 1946.

[71] H. Cramér, *Mathematical Methods of Statistics*. Princeton Univ. Press, Sep. 1946.

[72] P. J. Daniell, "Discussion of 'On the theoretical specification and sampling properties of autocorrelated time-series'," *J. Royal Stat. Soc., Series B*, vol. 8, pp. 88–90, 1946.

[73] M. Davy, "Multiple fundamental frequency estimation based on generative models," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. Springer, May 2006, ch. 7.

[74] M. Davy and S. J. Godsill, "Bayesian harmonic models for musical signal analysis," in *Bayesian Statistics VII*, J. Bernardo, J. Berger, A. Dawid, and A. Smith, Eds. Oxford University Press, 2002.

[75] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2498–2517, Apr. 2006.

[76] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[77] B. de Finetti, *Theory of Probability*. Wiley, 1974, first of two volumes translated from *Teoria Delle probabilità*, published 1970. The second volume appeared under the same title in 1975.

[78] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, Sep. 1999.

[79] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc., Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[80] Q. Ding and S. M. Kay, "Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1959–1969, May 2011.

[81] P. M. Djuric, "A model selection rule for sinusoids in white Gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, Jul. 1996.

[82] ——, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[83] P. M. Djuric and S. M. Kay, "Parameter estimation of chirp signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 12, pp. 2118–2126, Dec. 1990.

[84] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[85] L. Dou and R. J. W. Hodgson, "Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation I," *Inverse Problems*, vol. 11, no. 5, pp. 1069–1085, 1995.

[86] ——, "Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation II," *Inverse Problems*, vol. 12, no. 2, pp. 121–137, 1996.

[87] X. G. Doukopoulos and G. V. Moustakides, "Fast and stable subspace tracking," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1452–1465, Apr. 2008.

[88] M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," 2011, unpulished manuscript.

[89] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: A flexible Bayesian approach." *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1283–1295, 2007.

[90] B. Edler and H. Purnhagen, "Parametric audio coding," *Proc. Conf. Signal Process.*, vol. 1, pp. 21–24, Aug. 2000.

[91] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, 2nd ed. Springer, Jun. 1998.

[92] J. B. J. Fourier and G. Darboux, *Œuvres de Fourier*. Gauthier-Villars et Fils, 1890, no. 2.

[93] J. A. Foutz, A. Spanias, and M. K. Banavar, *Narrowband Direction of Arrival Estimation for Antenna Arrays*. Morgan & Claypool Publishers, Jul. 2008.

[94] D. Gabor, "Theory of communication," *J. Inst. Elect. Eng.*, vol. 93, no. 46, pp. 429–457, Nov. 1946.

[95] P. N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Commun.*, vol. 53, no. 8, pp. 991–1001, Oct. 2011.

[96] C. F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, 1809.

[97] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Jul. 2003.

[98] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, November 1984.

[99] W. R. Gilks, *Markov Chain Monte Carlo in Practice*. Chapman Hall/CRC, Dec. 1995.

[100] G.-O. Glentis, "A fast algorithm for APES and Capon spectral estimation," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4207–4220, Sep. 2008.

[101] S. J. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2002, pp. 1769–1772.

[102] ——, "Bayesian computational models for inharmonicity in musical instruments," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2005, pp. 283–286.

[103] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration*. Springer-Verlag, London, 1998.

[104] G. H. Golub and C. F. van Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, Oct. 1996.

[105] M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, 1997, pp. 2037–2040.

[106] M. M. Goodwin, *Adaptive Signal Models: Theory, Algorithms and Audio Applications*. Springer, Oct. 1998.

[107] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.

[108] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, Jan. 2003.

[109] C. Han and B. P. Carlin, "Markov chain Monte Carlo methods for computing Bayes factors: A comparative review," *J. Amer. Statistical Assoc.*, vol. 96, no. 455, pp. 1122–1132, Sep. 2001.

[110] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *J. Amer. Statistical Assoc.*, vol. 96, no. 454, pp. 746–774, 2001.

[111] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1989.

[112] A. C. Harvey and A. Jaeger, "Detrending, stylized facts and the business cycle," *J. Appl. Econometrics*, vol. 8, no. 3, pp. 231–47, Jul. 1993.

[113] A. Harvey, T. Trimbur, and H. v. Dijk, "Cyclical components in economic time series," Erasmus University Rotterdam, Econometric Institute, Econometric Institute Report EI 2002-20, Nov. 2002.

[114] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.

[115] S. Haykin, *Communication Systems*, 4th ed. John Wiley and Sons Ltd, May 2000.

[116] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. V. Huffel, "Perceptual audio modeling with exponentially damped sinusoids," *Signal Processing*, vol. 85, no. 1, pp. 163–176, Jan. 2005.

[117] U. Hoppe, F. Rosanowski, M. Döllinger, J. Lohscheller, M. Schuster, and U. Eysholdt, "Glissando: laryngeal motorics and acoustics," *J. Voice*, vol. 17, no. 3, pp. 370–376, Sep. 2003.

[118] D. V. Hoyt and K. H. Schatten, "New information on solar activity, 1779-1818, from Sir William Herschel's unpublished notebooks," *Astrophys. J.*, vol. 384, pp. 361–384, Jan. 1992.

[119] C. M. Hurvich and C.-L.Tsai, "A corrected Akaike information criterion for vector autoregressive model selection," *J. of Time Series Analysis*, vol. 14, no. 3, pp. 271–279, May 1993.

[120] A. Jakobsson and P. Stoica, "Combining Capon and APES for estimation of spectral lines," *Circ. Syst. Signal Process.*, vol. 19, no. 2, pp. 159–169, 2000.

[121] M. Jansson and P. Stoica, "Forward-only and forward-backward sample covariances – a comparative study," *Signal Processing*, vol. 77, no. 3, pp. 235–245, Sep. 1999.

[122] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, no. 3, pp. 227–241, 1968.

[123] ——, "Bayesian methods: general background," in *Max. Entropy and Bayesian Methods Appl. Stat.*, J. H. Justice, Ed., 1986.

[124] ——, "Bayesian spectrum and chirp analysis," in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G. J. Erickson, Eds. D. Reidel, Dordrecht-Holland, 1987, pp. 1–37.

[125] ——, *Probability Theory : The Logic of Science*, G. L. Bretthorst, Ed. Cambridge University Press, Apr. 2003.

[126] H. Jeffreys, *Theory of Probability*. Oxford University Press, 1939.

[127] J. Jensen, R. Heusdens, and S. H. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 121–132, Mar. 2004.

[128] J. H. Jensen, M. G. Christensen, and S. H. Jensen, "An amplitude and covariance matrix estimator for signals in colored gaussian noise," in *Proc. European Signal Processing Conf.*, 2009.

[129] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Pearson Prentice Hall, May 2008.

[130] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc. Interspeech*, 2005, pp. 537–540.

[131] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Prentice Hall, 1988.

[132] ——, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall PTR, Mar. 1993.

[133] S. M. Kay and S. L. Marple, Jr., "Spectrum analysis - a modern perspective," *Proc. IEEE*, vol. 69, no. 11, pp. 1380–1419, Nov. 1981.

[134] K. Kim and G. Shevlyakov, "Why Gaussianity?" *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 102–113, Mar. 2008.

[135] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[136] R. Kleijn and H. K. van Dijk, "Bayes model averaging of cyclical decompositions in economic time series," *Journal of Applied Econometrics*, vol. 21, no. 2, pp. 191–212, Mar. 2006.

[137] W. Kleijn and K. Paliwal, Eds., *Speech Coding and Synthesis*. Elsevier Science, Dec. 1995.

[138] J. Kovačević, V. K. Goyal, and M. Vetterli, *Signal Processing: Fourier and Wavelet Representations*. Cambridge University Press, Mar. 2012, version alpha 3.1 available. [Online]. Available: http://www.fourierandwavelets.org/

[139] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[140] R. Kumaresan and D. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 19, no. 1, pp. 134–139, Jan. 1983.

[141] S. Kunis, "Nonequispaced FFT - Generalisation and inversion," Ph.D. dissertation, Institut für Mathematik, Universität zu Lübeck, 2006.

[142] R. T. Lacoss, "Data adaptive spectral analysis methods," *Geophys.*, vol. 36, no. 4, pp. 661–675, Aug. 1971.

[143] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1625–1634, Jul. 2007.

[144] H. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Mathematica*, vol. 117, no. 1, pp. 37–52, 1967.

[145] P. S. Laplace, *Théorie Analytique des Probabilités*. Courcier, 1812.

[146] ——, "Memoir on the probability of the causes of events," *Statist. Sci.*, vol. 1, no. 3, pp. 364–378, Aug. 1986, english translation by S. M. Stigler.

[147] E. G. Larsson, J. Li, and P. Stoica, "High-resolution nonparametric spectral analysis: Theory and applications," in *High-resolution and robust signal processing*, Y. Hua, A. Gershman, and Q. Cheng, Eds. Marcel Dekker, Oct. 2003, ch. 4.

[148] E. G. Larsson and P. Stoica, "Fast implementation of two-dimensional APES and CAPON spectral estimators," *Multidimension. Syst. Signal Process.*, vol. 13, no. 1, pp. 35–53, 2002.

[149] A. M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

[150] H. Li, J. Li, and P. Stoica, "Performance analysis of forward-backward matched-filterbank spectral estimators," *IEEE Trans. Signal Process.*, vol. 46, no. 7, pp. 1954–1966, Jul. 1998.

[151] J. Li and P. Stoica, "An adaptive filtering approach to spectral estimation and SAR imaging," *IEEE Trans. Signal Process.*, vol. 44, no. 6, pp. 1469–1484, Jun. 1996.

[152] ——, "Efficient mixed-spectrum estimation with applications to target feature extraction," *IEEE Trans. Signal Process.*, vol. 44, no. 2, pp. 281–295, Feb. 1996.

[153] A. P. Liavas and P. A. Regalia, "On the behavior of information theoretic criteria for model order selection," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1689–1695, Aug. 2001.

[154] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 787–798, Sep. 2005.

[155] J. Lindblom and P. Hedelin, "Packet loss concealment based on sinusoidal modeling," in *IEEE Proc. Workshop on Speech Coding*, Oct. 2002, pp. 65–67.

[156] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, Jun. 2007.

[157] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Jun. 2002.

[158] S. Makino, T.-W. Lee, and H. Sawada, Eds., *Blind Speech Separation*. Springer, Sep. 2007.

[159] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Academic Press, Dec. 2008.

[160] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[161] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.

[162] S. L. Marple, Jr., *Digital Spectral Analysis: With Applications*. Prentice Hall, Jan. 1987.

[163] ——, "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.

[164] G. Marrelec, H. Benali, P. Ciuciu, M. Pélégrini-Issac, and J.-B. Poline, "Robust Bayesian estimation of the hemodynamic response function in event-related BOLD fMRI using basic physiological information," *Hum. Brain Mapp.*, vol. 19, no. 1, pp. 1–17, May 2003.

[165] F. Marvasti, *Nonuniform Sampling, Theory and Practice*. Springer, Jun. 2001.

[166] M. Mboup and T. Adalı, "A generalization of the fourier transform and its application to spectral analysis of chirp-like signals," *Appl. Comput. Harmon. Anal.*, vol. 32, no. 2, pp. 305–312, Mar. 2012.

[167] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[168] ——, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, vol. 495, ch. 4, pp. 121–173.

[169] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, Mar. 1953.

[170] F. P. Myburg, "Design of a scalable parametric audio coder," Ph.D. dissertation, Technische Universiteit Eindhoven, 2004.

[171] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, Oct. 1986.

[172] J. K. Nielsen, "Sinusoidal parameter estimation - a Bayesian approach," Master Thesis, Aalborg University, Denmark, Jun. 2009. [Online]. Available: http://kom.aau.dk/~jkn/publications/publications.php

[173] J. Nieuwenhuijse, R. Heusens, and E. F. Deprettere, "Robust exponential modeling of audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 6, May 1998, pp. 3581–3584.

[174] H. Nyquist, "Certain topics in telegraph transmission theory," *Proc. IEEE*, vol. 90, no. 2, pp. 280–305, Feb. 2002, reprint of the 1928 edition.

[175] J. N. Pandey, *The Hilbert Transform of Schwartz Distributions and Applications*. Wiley-Interscience, Feb. 1996.

[176] J.-M. Papy, L. De Lathauwer, and S. Van Huffel, "A shift invariance-based order-selection technique for exponential data modelling," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 473–476, Jul. 2007.

[177] A. Paulraj, R. Roy, and T. Kailath, "Estimation of signal parameters via rotational invariance techniques- ESPRIT," *Rec. Asilomar Conf. Signals, Systems, and Computers*, pp. 83–89, Nov. 1985.

[178] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophys. J. Roy. Astron. Soc.*, vol. 33, no. 3, pp. 347–366, Sep. 1973.

[179] C. J. Plack, A. J. Oxenham, R. R. Fay, and A. N. Popper, Eds., *Pitch: Neural Coding and Perception*. Springer, Aug. 2005.

[180] D. Potts, G. Steidl, and M. Tasche, "Modern sampling theory: Mathematics and applications," in *Modern Sampling Theory: Mathematics and Applications*, J. Benedetto and P. Ferreira, Eds. Birkhäuser Boston, 2001, ch. 12.

[181] H. Purnhagen and N. Meine, "HILN-the MPEG-4 parametric audio coding tools," *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 3, pp. 201–204, 2000.

[182] Y. Qi, T. P. Minka, and R. W. Picara, "Bayesian spectrum estimation of unevenly sampled nonstationary data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2002, pp. 1473–1476.

[183] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary, "AM-FM separation using auditory-motivated filters," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 465–480, Sep. 1997.

[184] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of Cal. Math. Soc.*, vol. 37, no. 3, pp. 81–91, 1945.

[185] C. R. Rao and Y. Wu, "On model selection," *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, vol. 38, pp. 1–57, 2001.

[186] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978.

[187] ——, "Estimation of structure by minimum description length," *Circuits, Systems, and Signal Process.*, vol. 1, no. 3, pp. 395–406, 1982.

[188] ——, "MDL denoising," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2537–2543, Nov. 2000.

[189] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer-Verlag New York, Inc., Jul. 2004.

[190] E. A. Robinson, "A historical perspective of spectrum estimation," *Proc. IEEE*, vol. 70, no. 9, pp. 885–907, Sep. 1982.

[191] C. Rodbro and S. Jensen, "Time-scaling of sinusoids for intelligent jitter buffer in packet based telephony," in *IEEE Proc. Workshop on Speech Coding*, Oct. 2002, pp. 71–73.

[192] T. Roos, P. Myllymaki, and J. Rissanen, "MDL denoising revisited," *IEEE Trans. Signal Process.*, vol. 57, no. 9, pp. 3347–3360, Sep. 2009.

[193] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*, 3rd ed. Academic Press, Jul. 2004.

[194] S. T. Roweis, "One microphone source separation," in *Adv. in Neural Inf. Process. Syst.*, 2000, pp. 793–799.

[195] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[196] D. V. Rubtsov and J. L. Griffin, "Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy," *J. Magnetic Resonance*, vol. 188, no. 2, pp. 367–379, Aug. 2007.

[197] C. A. Rødbro, M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Compressed domain packet loss concealment of sinusoidally coded speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2003, pp. 104–107.

[198] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1609–1623, Sep. 2006.

[199] D. F. Schmidt and E. Makalic, "The consistency of MDL for linear regression models with increasing signal-to-noise ratio," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1508–1510, Mar. 2012.

[200] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[201] A. Schuster, "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena," *Terr. Magn.*, vol. 3, no. 1, pp. 13–41, Mar. 1898.

[202] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[203] E. Serpedin, F. Panduru, I. Sarı, and G. B. Giannakis, "Bibliography on cyclostationarity," *Signal Processing*, vol. 85, no. 12, pp. 2233–2303, Dec. 2005.

[204] C. E. Shannon, "Communication in the presence of noise," *Proc. IEEE*, vol. 86, no. 2, pp. 447–457, Feb. 1998, reprint of the 1949 edition.

[205] SIDC-team, "The international sunspot number," Royal Observatory of Belgium, Ringlaan 3, 1180 Brussel, Belgium, 1818–2012. [Online]. Available: http://www.sidc.be/sunspot-data/

[206] D. S. Sivia, *Data Analysis: A Bayesian Tutorial*. Oxford University Press, Sep. 1996.

[207] R. J. Sluijter, "The development of speech coding and the first standard coder for public mobile telephony," Ph.D. dissertation, Techniche Universiteit Eindhoven, 2005.

[208] C. R. Smith, G. J. Erickson, and P. O. Neudorfer, "Parameter estimation in chirped signals," in *Proc. IEEE Conf. Commun., Comp., and Signal Process.*, Jun. 1989, pp. 538–539.

[209] W. W. Soon and S. H. Yaskell, *The Maunder Minimum and the variable sun-earth connection*. World Scientific Publishing Company, Aug. 2004.

[210] P. Stoica and P. Babu, "The Gaussian data assumption leads to the largest Cramér-Rao bound," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 132–133, May 2011.

[211] P. Stoica, A. Jakobsson, and J. Li, "Cisoid parameter estimation in the colored noise case: asymptotic Cramer-Rao bound, maximum likelihood, and nonlinear least-squares," *IEEE Trans. Signal Process.*, vol. 45, no. 8, pp. 2048–2059, Aug. 1997.

[212] ——, "Matched-filter bank interpretation of some spectral estimators," *Elsevier Signal Processing*, vol. 66, no. 1, pp. 45–59, 1998.

[213] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results, and an application," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 338–352, Feb. 2000.

[214] ——, "A new derivation of the apes filter," *IEEE Signal Process. Lett.*, vol. 6, no. 8, pp. 205–206, Aug. 1999.

[215] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Prentice Hall, May 2005.

[216] P. Stoica, R. L. Moses, B. Friedlander, and T. Söderström, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 378–392, Mar. 1989.

[217] P. Stoica and Y. Selén, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[218] P. Stoica, Y. Selén, and J. Li, "Multi-model approach to model selection," *Digital Signal Process.*, vol. 14, no. 5, pp. 399–412, Sep. 2004.

[219] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 7, pp. 1132–1143, Jul. 1990.

[220] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, vol. 495, ch. 14, pp. 495–518.

[221] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal," *J. Amer. Statistical Assoc.*, vol. 81, no. 393, pp. 82–86, Mar. 1986.

[222] H. L. V. Trees, *Optimum Array Processing*. Wiley-Interscience, Mar. 2002.

[223] R. E. Turner and M. Sahani, "Probabilistic amplitude and frequency demodulation," in *Adv. in Neural Inf. Process. Syst.*, 2011, pp. 981–989.

[224] ——, "Decomposing signals into a sum of amplitude and frequency modulated sinusoids using probabilistic inference." in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012.

[225] N. H. van Schijndel, J. Bensa, M. G. Christensen, C. Colomes, B. Edler, R. Heusdens, J. Jensen, S. H. Jensen, W. B. Kleijn, V. Kot, B. Kovesi, J. Lindblom, D. Massaloux, O. A. Niamut, F. Norden, J. H. Plasberg, R. Vafin, S. van de Par, D. Virette, and O. Wubbolt, "Adaptive RD optimized hybrid sound coding," *J. Audio Eng. Soc.*, vol. 56, no. 10, pp. 787–809, Oct. 2008.

[226] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[227] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, Mar. 2006.

[228] T. Verma and T. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Mar. 1999, pp. 981–984.

[229] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Process.*, vol. 39, no. 5, pp. 1110–1121, May 1991.

[230] E. Vincent, *MUSHRAM: A MATLAB interface for MUSHRA listening tests*, 2005. [Online]. Available: http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/

[231] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[232] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. 15, no. 2, pp. 70–73, Jun. 1967.

[233] W. Xu and M. Kaveh, "Analysis of the performance and sensitivity of eigendecomposition-based detectors," *IEEE Trans. Signal Process.*, vol. 43, no. 6, pp. 1413–1426, Jun. 1995.

[234] J. Yen, "On nonuniform sampling of bandwidth-limited signals," *IRE T. Circuit Theory*, vol. 3, no. 4, pp. 251–257, Dec. 1956.

[235] G. U. Yule, "On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers," *Phil. Trans. Roy. Soc. London, Series A*, vol. 226, pp. 267–298, Apr. 1927.

[236] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "A robust and computationally efficient subspace-based fundamental frequency estimator," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 487–497, mar. 2010.

Part II

Papers

Paper A

Bayesian Model Comparison with the g-Prior

Jesper Kjær Nielsen, Mads Græsbøll Christensen, Ali Taylan Cemgil, and Søren Holdt Jensen

The paper has been submitted to the
IEEE Transactions on Signal Processing,

In peer-review
The layout has been revised.

Abstract

Model comparison and selection is an important problem in many model-based signal processing applications. Often, very simple information criteria such as the Akaike information criterion or the Bayesian information criterion are used despite their shortcomings. Compared to these methods, Djuric's asymptotic MAP rule was an improvement, and in this paper we extend the work by Djuric in several ways. Specifically, we consider the elicitation of proper prior distributions, treat the case of real- and complex-valued data simultaneously in a Bayesian framework similar to that considered by Djuric, and develop new model selection rules for a regression model containing both linear and non-linear parameters. Moreover, we use this framework to give a new interpretation of the popular information criteria and relate their performance to the signal-to-noise ratio of the data. By use of simulations, we also demonstrate that our proposed model comparison and selection rules outperform the traditional information criteria both in terms of detecting the true model and in terms of predicting unobserved data.

1 Introduction

E*ssentially, all models are wrong, but some are useful* [1, p. 424]. This famous quote by Box accurately reflects the problem that scientists and engineers face when they analyse data originating from some physical process. As the exact description of a physical process is usually impossible due to the sheer amount of complexity or an incomplete knowledge, simplified and approximate models are often used instead. In this connection, model comparison and selection methods are vital tools for the elicitation of one or several models which can be used to make inference about physical quantities or to make predictions. Typical model selection problems are to find the number of non-zero regression parameters in linear regression [2–4], the number of sinusoids in a periodic signal [5–9], the orders of an autoregressive moving average (ARMA) process [10–15], and the number of clusters in a mixture model [16–18]. For several decades, a large variety of model comparison and selection methods have been developed (see, e.g., [3, 19–22] for an overview). These methods can basically be divided in three groups with the first group being those methods which require an a priori estimate of the model parameters, the second group being those methods which do not require such estimates, and the third group being those methods in which the model parameters and model are estimated and detected jointly [15]. The widely used information criteria such as the Akaike information criterion (AIC) [23], the corrected AIC ($\mathrm{AIC_c}$) [24], the generalised information criterion (GIC) [25], the Bayesian information criterion (BIC) [26], the minimum description length (MDL) [27, 28], the Hannan-Quinn information criterion (HQIC) [10], and the predictive least squares [29] belong to the first group of methods. The methods in the second group typically utilise a principal component analysis of the

data by analysing the eigenvalues [11, 15, 30], the eigenvectors [31, 32], or the angles between subspaces [33]. In the third group, the Bayesian methods are found. Although these methods are widely used in the statistical community [3, 34–37], their use in the signal processing community has only been limited (see, e.g., [7, 8, 14] for a few notable exceptions) compared to the use of the information criteria. The main reasons for this are the high computational costs of running these algorithms and the difficulty of specifying proper prior distributions. A few approximate methods have therefore been developed circumventing most of these issues. Two examples of such approximate methods are the BIC [26] and the asymptotic maximum a posteriori (MAP) rule [38, 39].

The BIC and the original MDL principle are equivalent, but they are derived using very different arguments [22, App. C]. Although this type of rule is one of the most popular model selection methods, it suffers from that every model parameter contributes with the same penalty to the overall model complexity penalty term in the model selection method. Djuric's asymptotic MAP rule [38] improves on the BIC method by accounting for that the magnitude of the penalty should depend on the type of models and model parameters being used. For example, the frequency parameter of a sinusoidal signal is shown to contribute with a three times larger penalty term than the sinusoidal amplitude and phase. Like the BIC method, the asymptotic MAP rule is derived in a Bayesian framework. However, in order to get very simple expressions, Djuric uses asymptotic considerations and improper priors, and he also neglects lower order terms during the derivations. The latter is a consequence of the use of improper priors.

In this paper, we extend the work by Djuric in several ways. First, we treat the difficult problem of eliciting proper and improper prior distributions on the model parameters. In this connection, we use a prior of the same form as the Zellner's g-prior [40], discuss its properties, and re-parametrise it in terms of the signal-to-noise ratio (SNR) to facilitate a better understanding of it. Second, we treat real- and complex-valued signals simultaneously and propose a few new model selection rules, and third, we derive the most common information criteria in our framework. The latter is useful for assessing the conditions under which the, e.g., AIC and BIC are accurate. As opposed to the various information criteria which are generally derived from cross-validation using the Kullback-Leibler (KL) divergence, we analyse the model comparison problem in a Bayesian framework for numerous reasons [34, 35]; Bayesian model comparison is consistent under very mild conditions, naturally selects the simplest model which explains the data reasonably well (the principle of Occam's razor), takes model uncertainty into account for estimation and prediction, works for non-nested models, enables a more intuitive interpretation of the results, and is conceptually the same, regardless of the number and types of models under consideration. The two major disadvantages of Bayesian model comparison are that the computational cost of running the resulting algorithms may be too high, and that the use of improper and vague prior distributions only leads to sensible answers under certain circumstances. In this paper, we discuss and address both of these issues.

The paper is organised as follows. In Sec. 2, we give an introduction to model comparison in a Bayesian framework and discuss some of the difficulties associated with the elicitation of prior distributions and the evaluation of the marginal likelihood. In Sec. 3, we propose a general regression model consisting of both linear- and non-linear parameters. For known non-linear parameters, we derive two model comparison algorithms in Sec. 4 and give a new interpretation of the traditional information criteria. For unknown non-linear parameters, we also derive two model comparison algorithms rules in Sec. 5. Through simulations, we evaluate the proposed model comparison algorithms in Sec. 6, and Sec. 7 concludes this paper.

2 Bayesian Model Comparison

We assume that we observe some real- or complex-valued data $\{x(t_n)\}_{n=0}^{N-1}$ which we write as the vector

$$\boldsymbol{x} = \begin{bmatrix} x(t_0) & x(t_1) & \cdots & x(t_{N-1}) \end{bmatrix}^T , \tag{A.1}$$

and we assume that these data originate from some unknown model. Since we are unsure about the true model, we select a set of K candidate parametric models $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_K$ which we wish to compare in the light of the data $\boldsymbol{x}$. Each model $\mathcal{M}_k$ is parametrised by the model parameters $\boldsymbol{\theta}_k \in \Theta_k$ where Θ_k is the parameter space of dimension d_k. The relationship between the data $\boldsymbol{x}$ and the model $\mathcal{M}_k$ is given by the probability distribution with density[1] $p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)$ which is called the observation model. When viewed as a function of the model parameters, the observation model is referred to as the likelihood function. The likelihood function plays an important role in statistics where it is used for parameter estimation. However, model selection cannot be solely based on comparing candidate models in terms of their likelihood as a complex model can be made to fit the observed data better than a simple model. The various information criteria are alternative ways of resolving this by introducing a term that penalizes more complex models. This is a manifestation of the well known *Occam's razor* principle which states that if two models explain the data equally well, the simplest model should always be preferred [41, p. 343].

In a Bayesian framework, the model parameters and the model are random variables with the pdf $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ and pmf $p(\mathcal{M}_k)$, respectively. We refer to these distributions as the prior distributions as they contain our state of knowledge before any data are observed. After observing data, we update our state of knowledge by transforming the prior distributions into the posterior pdfs $p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k)$ and pmf $p(\mathcal{M}_k|\boldsymbol{x})$. The prior and posterior distributions for the model parameters and the model are connected by

[1] In this paper, we have used the generic notation $p(\cdot)$ to denote both a probability density function (pdf) over a continuous parameter and a probability mass function (pmf) over a discrete parameter.

Bayes' theorem

$$p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{p(\boldsymbol{x}|\mathcal{M}_k)} \tag{A.2}$$

$$p(\mathcal{M}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{M}_k)p(\mathcal{M}_k)}{p(\boldsymbol{x})} \tag{A.3}$$

where

$$p(\boldsymbol{x}|\mathcal{M}_k) = \int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k \tag{A.4}$$

is called the marginal likelihood or the evidence. For model comparison, we often compare the odds of two competing models $\mathcal{M}_j$ and $\mathcal{M}_i$. In this connection, we define the posterior odds which are given by

$$\frac{p(\mathcal{M}_j|\boldsymbol{x})}{p(\mathcal{M}_i|\boldsymbol{x})} = \mathrm{BF}[\mathcal{M}_j; \mathcal{M}_i]\frac{p(\mathcal{M}_j)}{p(\mathcal{M}_i)} \tag{A.5}$$

where the Bayes' factor is given by

$$\mathrm{BF}[\mathcal{M}_j; \mathcal{M}_i] = \frac{p(\boldsymbol{x}|\mathcal{M}_j)}{p(\boldsymbol{x}|\mathcal{M}_i)} \triangleq \frac{m_j(\boldsymbol{x})}{m_i(\boldsymbol{x})} \tag{A.6}$$

where $m_k(\boldsymbol{x})$ is an unnormalised marginal likelihood whose normalisation constant must be the same for all models. Working with $m_k(\boldsymbol{x})$ rather than the normalised marginal likelihood $p(\boldsymbol{x}|\mathcal{M}_k)$ is usually much simpler. Moreover, $p(\boldsymbol{x}|\mathcal{M}_k)$ does not even exist if improper priors are used. We return to this in Sec 2.1. Since the prior and posterior distributions of the model are discrete, it is easy to find the posterior odds and the posterior distribution once the Bayes' factors are known. For example, we may rewrite the posterior distribution for the models in terms of the Bayes' factors as

$$p(\mathcal{M}_k|\boldsymbol{x}) = \frac{\mathrm{BF}[\mathcal{M}_k; \mathcal{M}_b]p(\mathcal{M}_k)}{\sum_{i=1}^{K} \mathrm{BF}[\mathcal{M}_i; \mathcal{M}_b]p(\mathcal{M}_i)} \tag{A.7}$$

where $\mathcal{M}_b$ is some user selected base model which all other models are compared against. Therefore, the main computational challenge in Bayesian model comparison is to compute the unnormalised marginal likelihoods, constituting the Bayes' factor for competing pairs of models. We return to this in Sec 2.2. The posterior distribution on the models may be used to select the most probable model. However, as the posterior distribution contains the probabilities of all candidate models, all models may be used to make inference about the unknown parameters or to predict unobserved data points. This is called Bayesian model averaging. For example, assume that we are interested in predicting a future data vector $\boldsymbol{x}_p$ using all models. The predictive distribution then has the density

$$p(\boldsymbol{x}_p|\boldsymbol{x}) = \sum_{k=1}^{K} p(\mathcal{M}_k|\boldsymbol{x})p(\boldsymbol{x}_p|\boldsymbol{x}, \mathcal{M}_k) \ . \tag{A.8}$$

Thus, the model averaged prediction is a weighted sum of the predictions from every model.

2.1 On the Use of Improper Prior Distributions

Like Djuric [38, 39], we might be tempted to use improper prior distributions when we have no or little prior information before observing any data. Whereas this usually works for the inference about model parameters, it usually leads to indeterminate Bayes' factors. To see this, let the prior distribution on the model parameters of the k'th model have the joint density $p(\boldsymbol{\theta}_k|\mathcal{M}_k) = c_k^{-1}h(\boldsymbol{\theta}_k|\mathcal{M}_k)$ where $c_k = \int_{\Theta_k} h(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k$ is the normalisation constant. In the limit $c_k \to \infty$, the prior distribution is said to be improper. An example of a popular improper prior pdf is $h(\boldsymbol{\theta}_k|\mathcal{M}_k) = 1$ so that $p(\boldsymbol{\theta}_k|\mathcal{M}_k) \propto 1$ where $\propto$ denotes proportional to. The posterior distribution on the model parameters has the pdf

$$p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{p(\boldsymbol{x}|\mathcal{M}_k)} \tag{A.9}$$

$$= \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)h(\boldsymbol{\theta}_k|\mathcal{M}_k)}{\int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)h(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k} \,. \tag{A.10}$$

Thus, provided that the integral

$$\tilde{p}(\boldsymbol{x}|\mathcal{M}_k) = \int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)h(\boldsymbol{\theta}_k|\mathcal{M}_k)d\boldsymbol{\theta}_k \tag{A.11}$$

converges, the posterior pdf $p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k)$ is proper even for an improper prior distribution. For two competing models $\mathcal{M}_j$ and $\mathcal{M}_i$, the Bayes' factor is

$$\mathrm{BF}[\mathcal{M}_j; \mathcal{M}_i] = \frac{c_i}{c_j}\frac{\tilde{p}(\boldsymbol{x}|\mathcal{M}_j)}{\tilde{p}(\boldsymbol{x}|\mathcal{M}_i)} \,. \tag{A.12}$$

The ratio $\tilde{p}(\boldsymbol{x}|\mathcal{M}_j)/\tilde{p}(\boldsymbol{x}|\mathcal{M}_i)$ is well-defined if the posterior distributions on the model parameters $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_i$ are proper. For proper prior distributions, the scalars c_i and c_j are finite, and the Bayes' factor is therefore well-defined. However, for improper prior distributions, the Bayes' factor is in general indeterminate. Specifically, for the popular improper prior distribution with $h(\boldsymbol{\theta}_j|\mathcal{M}_j) = h(\boldsymbol{\theta}_i|\mathcal{M}_i) = 1$, it can be shown that [42]

$$\frac{c_i}{c_j} = \begin{cases} 0 \,, & d_j > d_i \\ 1 \,, & d_j = d_i \\ \infty \,, & d_j < d_i \end{cases} \tag{A.13}$$

where d_j and d_i are the number of model parameters in $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_i$, respectively. That is, the simplest model is always preferred over more complex models, regardless of the

information in the data. This phenomenon is known as the *Bartlett's paradox*[2] [43]. Due to the Bartlett's paradox, the general recommendation is that one should use proper prior distributions for model comparison. However, there exists one important exception to this rule which we consider below. From (A.12), we also see that vague prior distributions may give misleading answers. For example, a vague distribution such as the normal distribution with a very large variance leads to an arbitrary large normalising constant c_k which strongly influences the Bayes' factor [35]. Therefore, the elicitation of proper prior distributions is very important for Bayesian model comparison.

Common Model Parameters

Consider the case where one model, the null model $\mathcal{M}_N$, is a sub-model[3] of all other candidate models. That is $\mathcal{M}_N \subseteq \mathcal{M}_k$ for $k = 1, \ldots, K$. We denote the null model parameters as $\boldsymbol{\theta}_N$ and the model parameters of the k'th model as $\boldsymbol{\theta}_k = \begin{bmatrix} \boldsymbol{\theta}_N^T & \boldsymbol{\psi}_k^T \end{bmatrix}^T$ where $(\cdot)^T$ denotes matrix transposition. The prior distribution on $\boldsymbol{\theta}_k$ now has the pdf

$$p(\boldsymbol{\theta}_k|\mathcal{M}_k) = p(\boldsymbol{\psi}_k|\boldsymbol{\theta}_N, \mathcal{M}_k)p(\boldsymbol{\theta}_N|\mathcal{M}_k) \ . \tag{A.14}$$

If the null model parameters have the same meaning[4] in $\mathcal{M}_k$ and $\mathcal{M}_N$, then $p(\boldsymbol{\theta}_N|\mathcal{M}_k) = p(\boldsymbol{\theta}_N|\mathcal{M}_N)$. Thus, using the prior pdf $p(\boldsymbol{\theta}_N|\mathcal{M}_N) = c_b^{-1}h(\boldsymbol{\theta}_N|\mathcal{M}_N)$, the Bayes' factor is

$$\mathrm{BF}[\mathcal{M}_k; \mathcal{M}_N] = \frac{\int_{\Theta_k} p(\boldsymbol{x}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\psi}_k|\boldsymbol{\theta}_N, \mathcal{M}_k)h(\boldsymbol{\theta}_N|\mathcal{M}_N)d\boldsymbol{\theta}_k}{\int_{\Theta_b} p(\boldsymbol{x}|\boldsymbol{\theta}_N, \mathcal{M}_N)h(\boldsymbol{\theta}_N|\mathcal{M}_N)d\boldsymbol{\theta}_N} \tag{A.16}$$

which is proper if the posterior distribution on the null model parameters and the prior distribution with pdf $p(\boldsymbol{\psi}_k|\boldsymbol{\theta}_N, \mathcal{M}_k)$ are proper. That is, the Bayes' factor is well-defined since $c_b = c_i = c_j$ even if an improper prior distribution is selected on the null model parameters, provided that they have the same meaning across all candidate models.

2.2 Computing the Marginal Likelihood

As alluded to earlier, the main computational difficulty in computing the posterior distribution on the models is the evaluation of the marginal likelihood in (A.4). The integral may not have a closed-form solution, and direct numerical evaluation may be

[2] Bartlett's paradox is also called the Lindley's paradox, the Jeffreys' paradox, and various combinations of the three names.

[3] Instead of the null model, the full model, which contains all other candidate models, can also be used [4].

[4] If one set of parameters $\boldsymbol{\theta}_N$ has the same meaning in two nested models $\mathcal{M}_N$ and $\mathcal{M}_k$ with $\mathcal{M}_N \subset \mathcal{M}_k$, the Fisher information matrix of the model parameters $\boldsymbol{\theta}_k$ is diagonal. That is,

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_k) = \boldsymbol{\mathcal{I}}(\boldsymbol{\psi}_k, \boldsymbol{\theta}_N) = \begin{bmatrix} \boldsymbol{\mathcal{I}}(\boldsymbol{\psi}_k) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_N) \end{bmatrix} \ . \tag{A.15}$$

infeasible if the number of model parameters is too large. Numerous solutions to this problem have been proposed and they can broadly be categorised as stochastic methods and deterministic methods. In the stochastic methods, the integral is evaluated using numerical sampling which are also known as Monte Carlo techniques [44]. Popular techniques are importance sampling [45], Chib's methods [46, 47], and reversible jump Markov chain Monte Carlo [48]. An overview over and comparison of several methods are given in [49]. An advantage of the stochastic methods is that they in principle can generate exact results. However, it might be difficult to assess the convergence of the underlying stochastic integration algorithm. On the other hand, the deterministic methods can only generate approximate results since they are based on analytical approximations which make the evaluation of the integral in (A.4) possible. These methods are also sometimes referred to as variational Bayesian methods [50], and a simple and widely used example of these methods is the Laplace approximation [51]. In order to derive the BIC and the MAP rule and since the Laplace approximation is used later in this paper, we briefly describe it here.

The Laplace Approximation

Denote the integrand of an integral such as in (A.4) by $f(\boldsymbol{\xi}_k)$ where $\boldsymbol{\xi}_k = \begin{bmatrix}\mathrm{Re}(\boldsymbol{\theta}_k^T) & \mathrm{Im}(\boldsymbol{\theta}_k^T)\end{bmatrix}^T$ is a vector of $\tilde{d}_k$ real parameters with support Ξ_k. Moreover, suppose there exists a suitable one-to-one transformation $\boldsymbol{\xi}_k = \boldsymbol{h}(\boldsymbol{\varphi}_k)$ such that the logarithm of the integrand

$$q(\boldsymbol{\varphi}_k) = \left|\frac{\partial \boldsymbol{h}(\boldsymbol{\varphi}_k)}{\boldsymbol{\varphi}_k}\right| f\left(\boldsymbol{h}(\boldsymbol{\varphi}_k)\right) \tag{A.17}$$

can be accurately approximated by the second-order Taylor expansion around a mode $\hat{\boldsymbol{\varphi}}_k$ of $q(\boldsymbol{\varphi}_k)$. That is,

$$\ln q(\boldsymbol{\varphi}_k) \approx q(\hat{\boldsymbol{\varphi}}_k) + \frac{1}{2}(\boldsymbol{\varphi}_k - \hat{\boldsymbol{\varphi}}_k)^T \boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)(\boldsymbol{\varphi}_k - \hat{\boldsymbol{\varphi}}_k) \tag{A.18}$$

where

$$\boldsymbol{H}(\boldsymbol{\varphi}_k) = \frac{\partial^2 \ln q(\boldsymbol{\varphi}_k)}{\partial \boldsymbol{\varphi}_k \partial \boldsymbol{\varphi}_k^T} \tag{A.19}$$

is the Hessian matrix. Under certain regularity conditions [39], the Laplace approximation is then given by

$$\int_{\Phi_k} q(\boldsymbol{\varphi}_k) d\boldsymbol{\varphi}_k \approx q(\hat{\boldsymbol{\varphi}}_k)(2\pi)^{\tilde{d}_k/2} |-\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)|^{-1/2} \tag{A.20}$$

where Φ_k is the support of $\boldsymbol{\varphi}_k$. The main difficulty in computing the Laplace approximation is to find a suitable parametrisation of the integrand so that the second-order Taylor expansion of $\ln q(\boldsymbol{\varphi}_k)$ is accurate. If $q(\boldsymbol{\varphi}_k)$ consists of multiple, significant, and well-separated peaks, an integral can be approximated by a Laplace approximation to each peak at their respective modes [52].

BIC and Asymptotic MAP

The BIC [26] and the asymptotic MAP rule [39] are based on the Laplace approximation with $\boldsymbol{h}(\cdot)$ being the identity function so that

$$f(\boldsymbol{\xi}_k) = q(\boldsymbol{\varphi}_k) = p(\boldsymbol{x}|\boldsymbol{\xi}_k, \mathcal{M}_k)p(\boldsymbol{\xi}_k|\mathcal{M}_k) \ . \tag{A.21}$$

By neglecting terms of order $\mathcal{O}(1)$ and assuming a flat prior around $\hat{\boldsymbol{\xi}}_k$, the marginal likelihood in the asymptotic MAP rule is

$$\int_{\Xi_k} f(\boldsymbol{\xi}_k) d\boldsymbol{\xi}_k \approx p(\boldsymbol{x}|\hat{\boldsymbol{\xi}}_k, \mathcal{M}_k)|-\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)|^{-1/2} \ . \tag{A.22}$$

In the MAP rule, the determinant of the observed information matrix $-\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)$ is evaluated using asymptotic considerations, and the asymptotic result therefore depends on the specific structure of $\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)$. For the BIC, however, this determinant is assumed to grow linearly in the sample size N so that

$$|-\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)| = \left|-\frac{N}{\alpha}\frac{\alpha}{N}\boldsymbol{H}(\hat{\boldsymbol{\varphi}}_k)\right| = \left(\frac{N}{\alpha}\right)^{\tilde{d}_k} \mathcal{O}(1) \tag{A.23}$$

where α is an arbitrary constant. In the BIC, $\alpha = 1$ and the BIC is therefore

$$\int_{\Xi_k} f(\boldsymbol{\xi}_k) d\boldsymbol{\xi}_k \approx p(\boldsymbol{x}|\hat{\boldsymbol{\xi}}_k, \mathcal{M}_k) N^{-\tilde{d}_k/2} \ , \tag{A.24}$$

but α can be selected arbitrarily which we find unsatisfactory. In [39], Djuric shows that the MAP rule and the BIC coincide for autoregressive models and sinusoidal models with known frequencies. However, he also shows that that they differ for polynomial models, sinusoidal models with unknown frequencies, and chirped signal models.

3 Model Comparison in Regression Models

Bayesian model comparison as outlined in Sec. 2 is applicable to any model, but we have to work with a specific model to come up with specific algorithms for model comparison. In the rest of this paper, we therefore focus on regression models of the form

$$\mathcal{M}_k : \ \boldsymbol{x} = \boldsymbol{s}_k(\boldsymbol{\phi}_k, \boldsymbol{\psi}, \boldsymbol{\alpha}_k) + \boldsymbol{e} = \boldsymbol{B}\boldsymbol{\psi} + \boldsymbol{Z}_k(\boldsymbol{\phi}_k)\boldsymbol{\alpha}_k + \boldsymbol{e} \tag{A.25}$$

where $\boldsymbol{s}_k(\boldsymbol{\phi}_k, \boldsymbol{\psi}, \boldsymbol{\alpha}_k)$ and $\boldsymbol{e}$ form a Wold decomposition of the real- or complex-valued data $\boldsymbol{x}$ into a predictable part and a non-predictable part, respectively. Since the model parameters are treated as random variables, the predictable part $\boldsymbol{s}_k(\boldsymbol{\phi}_k, \boldsymbol{\psi}, \boldsymbol{\alpha}_k)$ is also stochastic like the non-predictable part. All models include the same null model

$$\mathcal{M}_N : \ \boldsymbol{x} = \boldsymbol{B}\boldsymbol{\psi} + \boldsymbol{e} \tag{A.26}$$

where $\boldsymbol{B}$ and $\boldsymbol{\psi}$ are a known $N \times l_N$ system matrix and a known or unknown vector of l_N linear parameters, respectively. Usually, the predictable part of the null model is either taken to be a vector of ones so that $\boldsymbol{\psi}$ acts as an intercept or not present at all. In the latter case, the null model is simply the noise-only model. The various candidate models differ in terms of the l_k linear parameters in the vector $\boldsymbol{\alpha}_k$ and the $N \times l_k$ system matrix $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$, which is parametrised by the ρ_k real-valued and non-linear parameters in the vector $\boldsymbol{\phi}_k$. These non-linear parameters may be either known, unknown or not present at all. We discuss the first and latter case in Sec. 4 and the case of unknown non-linear parameters in Sec. 5. Without loss of generality, we assume that the columns of $\boldsymbol{B}$ and $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$ are orthogonal to each other so that $\boldsymbol{\psi}$ has the same interpretation in all models and therefore can be assigned an improper prior if $\boldsymbol{\psi}$ is unknown. If the columns of $\boldsymbol{B}$ and $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$ are not orthogonal to each other, $\boldsymbol{s}(\boldsymbol{\phi}_k, \boldsymbol{\psi}, \boldsymbol{\alpha}_k)$ can be re-parametrised so that the columns of the two system matrices are orthogonal [53]. We focus on the regression model in (A.25) for several reasons. First of all, many common signal models used in signal processing can be written in the form of (A.25). Examples of such models are the linear regression model, the polynomial regression model, the autoregressive signal model, the sinusoidal model, and the chirped signal model, and these five signal models were also considered by Djuric in [39]. Second, the regression model in (A.25) is analytically tractable and therefore results in computational algorithms with a tractable complexity. Moreover, the analytical tractability facilitates insight into, e.g., the various information criteria. Finally, the regression model in (A.25) can be viewed as a simple approximation to more complex models [3].

3.1 Elicitation of Prior Distributions

In the Bayesian framework, the unknown parameters are random variables. In addition to specifying a distribution on the noise vector, we therefore also have to elicit prior distributions on these unknown parameters. The elicitation of prior distributions is a controversial aspect in Bayesian statistics as it is often argued that subjectivity is introduced into the analysis. We here take a more practical view at this philosophical problem and consider the elicitation as a consistent and explicit way of stating our assumptions. In addition to the philosophical issue, we also face two practical problems in the context of eliciting prior distributions for model comparison. First, if we assume that $l_k \leq L$, we can select a subset of columns from $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$ in $K = 2^L$ different ways. A careful elicitation of the prior distribution for the model parameters in each model is therefore infeasible if L is too large, and we therefore prefer to do the elicitation in a more generic way. Second, even if we have only a vague prior knowledge, the use of improper or vague prior distributions in an attempt to be objective may lead to bad or non-sensible answers [35]. As we discussed in Sec. 2, this approach usually works for making inference about model parameters, but may lead to the Bartlett's paradox for model selection.

The Noise Distribution

In order to deduce the observation model, we have to select a model for the non-predictable part $\boldsymbol{e}$ of the model in (A.25). As it is purely stochastic, it must have zero mean, and we assume that it has a finite variance. As advocated by Jaynes and Bretthorst [54–56], we select the distribution which maximises the entropy under these constraints. It is well-known, that this distribution is the (complex) normal distribution with pdf

$$p(\boldsymbol{e}|\sigma^2) = \left[r\pi\sigma^2\right]^{-N/r} \exp\left(-\frac{\boldsymbol{e}^H\boldsymbol{e}}{r\sigma^2}\right) \tag{A.27}$$

$$= \begin{cases} \mathcal{CN}(\boldsymbol{e}; \boldsymbol{0}, \sigma^2\boldsymbol{I}_N) \ , & r = 1 \\ \mathcal{N}(\boldsymbol{e}; \boldsymbol{0}, \sigma^2\boldsymbol{I}_N) \ , & r = 2 \end{cases} \tag{A.28}$$

where $(\cdot)^H$ denotes conjugate matrix transposition, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and r is either 1 if $\boldsymbol{x}$ is complex-valued or 2 if $\boldsymbol{x}$ is real-valued. To simplify the notation, we use the non-standard notation $\mathcal{N}_r(\cdot)$ to refer to either the complex normal distribution with pdf $\mathcal{CN}(\cdot)$ for $r = 1$ or the real normal distribution with pdf $\mathcal{N}(\cdot)$ for $r = 2$. It is important to note that the noise variance σ^2 is a random variable. As opposed to the case where it is simply a fixed but unknown quantity, the noise distribution marginalised over this random noise variance is able to model noise with heavy tails and is robust towards outliers. Another important observation is that (A.28) does not explicitly model any correlations in the noise. However, including correlation constraints into the elicitation of the noise distribution lowers the entropy of the noise distribution which is therefore more informative [55, Ch. 7], [56]. This leads to more accurate estimates when there is genuine prior information about the correlation structure. However, if nothing is known about the correlation structure, the noise distribution in (A.28) is the best choice since it is the least informative distribution and is thus able to capture every possible correlation structure in the noise [56, 57].

The Gaussian assumption on the noise implies that the observed data are distributed as

$$p(\boldsymbol{x}|\boldsymbol{\alpha}_k, \boldsymbol{\psi}, \boldsymbol{\phi}_k, \sigma^2, \mathcal{M}_k) = \mathcal{N}_r(\boldsymbol{x}; \boldsymbol{B}\boldsymbol{\psi} + \boldsymbol{Z}_k(\boldsymbol{\phi}_k)\boldsymbol{\alpha}_k, \sigma^2\boldsymbol{I}_N) \ . \tag{A.29}$$

The Fisher information matrix (FIM) for this observation model is derived in App. A and given by (A.77). The block diagonal structure of the FIM means that the common parameters $\boldsymbol{\psi}$ and σ^2 have the same meaning in all models and can therefore be assigned improper prior distributions.

The Noise Variance

The noise variance is a common parameter and has the same meaning in all models, and it can therefore be assigned an improper prior. The Jeffreys' prior $p(\sigma^2) = (\sigma^2)^{-1}$

is a widely used improper prior for the noise variance which we also adopt in this paper. The popularity primarily stems from that the prior is invariant under transformations of the form σ^m for all $m \neq 0$. Thus, the Jeffreys' prior includes the same prior knowledge whether we parametrise our model in terms of the noise variance σ^2, the standard deviation σ, or the precision parameter $\lambda = \sigma^{-2}$.

The Linear Parameters

Since we have assumed that $\boldsymbol{B}^H \boldsymbol{Z}_k(\boldsymbol{\phi}_k) = \boldsymbol{0}$, the linear parameters $\boldsymbol{\psi}$ of the null model have the same meaning in all models. We can therefore use the improper prior distribution with pdf $p(\boldsymbol{\psi}) \propto 1$ for $\boldsymbol{\psi}$. This prior is often used for location parameters as it is translation invariant. As the dimension of the vector $\boldsymbol{\alpha}_k$ of linear parameters varies between models, a proper prior distribution must be assigned on it. For linear regression models, the Zellner's g-prior given by [40]

$$p(\boldsymbol{\alpha}_k | g, \sigma^2, \boldsymbol{\phi}_k, \mathcal{M}_k) = \mathcal{N}_r(\boldsymbol{\alpha}_k; \boldsymbol{0}, g\sigma^2 [\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k) \boldsymbol{Z}_k(\boldsymbol{\phi}_k)]^{-1}) \tag{A.30}$$

has been widely adopted since it leads to analytically tractable marginal likelihoods and is easy to understand and interpret [4]. The g-prior can be interpreted as the posterior distribution on $\boldsymbol{\alpha}_k$ arising from the analysis of a conceptual sample $\boldsymbol{x}_0 = \boldsymbol{0}$ given the non-linear parameters $\boldsymbol{\phi}_k$, a uniform prior on $\boldsymbol{\alpha}_k$, and a scaled variance $g\sigma^2$ [58]. Given $\boldsymbol{\phi}_k$, the covariance matrix of the g-prior also coincides with a scaled version of the inverse Fisher information matrix. Consequently, a large prior variance is therefore assigned to parameters which are difficult to estimate. We can also make a physical interpretation of the scalar g when the null model is the noise-only model. In this case, the average signal-to-noise ratio (SNR) of the data is

$$\eta = \frac{E\{\boldsymbol{\alpha}_k^H \boldsymbol{Z}_k^H(\boldsymbol{\phi}_k) \boldsymbol{Z}_k(\boldsymbol{\phi}_k) \boldsymbol{\alpha}_k\}}{E\{\boldsymbol{e}^H \boldsymbol{e}\}} \tag{A.31}$$

$$= \frac{E\{\mathrm{tr}[\boldsymbol{Z}_k(\boldsymbol{\phi}_k) E\{\boldsymbol{\alpha}_k \boldsymbol{\alpha}_k^H\} \boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)]\}}{N\sigma^2} = \frac{g l_k}{N} \tag{A.32}$$

where $E\{\cdot\}$ and $\mathrm{tr}(\cdot)$ denote the statistical expectation operator and the matrix trace, respectively. Thus, the value of g has a very simple relationship to the average SNR when the null model is the noise-only model.

If the hyperparameter g are treated as a fixed but unknown quantity, its value must be selected carefully. In, e.g., [2, 4, 59], the consequences of selecting various fixed choices of g have been analysed and evaluated, and as we also show in Sec. 4, the popular information criteria such as the AIC and BIC can be viewed as different ways of selecting a particular value of g. In [4, 60], the hyperparameter g was also treated as a random variable and integrated out of the marginal likelihood, thus avoiding the selection of a particular value for it. For the prior distribution on g, a special case of

the beta prime or inverted beta distribution with pdf

$$p(g|\delta) = \frac{\delta - r}{r}(1+g)^{-\delta/r} \ , \qquad \delta > r \ . \tag{A.33}$$

was used. The hyperparameter δ should be selected in the interval $r < \delta \leq 2r$ [4]. Besides having some desirable analytical properties, $p(g|\delta)$ reduces to the Jeffreys' prior and the reference prior for a linear regression model when $\delta = r$ [61]. However, since this prior is improper, it can only be used when the prior probability of the null model is zero.

The non-linear Parameters

The elicitation of the prior distribution on the non-linear parameters $\boldsymbol{\phi}_k$ is hard to do in general. In this paper, we therefore treat the case of fixed but unknown non-linear parameters and the case of non-linear parameters with a uniform prior of the form

$$p(\boldsymbol{\phi}_k|\mathcal{M}_k) = W_k^{-1}\mathbb{I}_{\Phi_k}(\boldsymbol{\phi}_k) \tag{A.34}$$

where $\mathbb{I}_{\Phi_k}(\cdot)$ is the indicator function on the support Φ_k. This uniform prior is often used for the non-linear parameters of sinusoidal and chirped signal models.

The Models

For the prior on the model, we select a uniform prior of the form $p(\mathcal{M}_k) = K^{-1}\mathbb{I}_{\mathcal{K}}(k)$ where $\mathcal{K} = \{1, 2, \ldots, K\}$. For a finite number of models, however, it is easy to use a different prior in our framework through (A.7).

3.2 Bayesian Inference

So far, we have elicited our probability model consisting of the observation model in (A.29) and the prior distributions on the model parameters. These distributions constitute the integrand of the integral representation of the marginal likelihood in (A.4), and we now evaluate this integral. After some algebra, the integrand can be rewritten as

$$\begin{aligned}
&p(\boldsymbol{x}|\boldsymbol{\alpha}_k, \boldsymbol{\psi}, \boldsymbol{\phi}_k, \sigma^2, \mathcal{M}_k)p(\boldsymbol{\alpha}_k|g, \sigma^2, \boldsymbol{\phi}_k, \mathcal{M}_k) \\
&\quad \times p(\boldsymbol{\psi})p(\sigma^2)p(g|\delta)p(\boldsymbol{\phi}_k|\mathcal{M}_k) \\
&\propto \mathcal{N}_r(\boldsymbol{\alpha}; c\boldsymbol{m}_k, g\sigma^2[\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k)]^{-1}/(1+g)) \\
&\quad \times \mathcal{N}_r(\boldsymbol{\psi}; \boldsymbol{m}_N, \sigma^2[\boldsymbol{B}^H\boldsymbol{B}]^{-1})\text{Inv-}\mathcal{G}\left(\sigma^2; \frac{N - l_N}{r}, \frac{N\hat{\sigma}_k^2}{r}\right) \\
&\quad \times m_N(\boldsymbol{x})(1+g)^{-l_k/r}\left(\frac{\hat{\sigma}_N^2}{\hat{\sigma}_k^2}\right)^{(N-l_N)/r} p(g|\delta)p(\boldsymbol{\phi}_k|\mathcal{M}_k)
\end{aligned} \tag{A.35}$$

where Inv-$\mathcal{G}$ is the inverse gamma distribution. Moreover, we have defined

$$\boldsymbol{m}_k \triangleq [\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k)]^{-1}\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{x} \tag{A.36}$$

$$\boldsymbol{m}_N \triangleq (\boldsymbol{B}^H\boldsymbol{B})^{-1}\boldsymbol{B}^H\boldsymbol{x} \tag{A.37}$$

$$\hat{\sigma}_k^2 \triangleq \frac{\boldsymbol{x}^H(\boldsymbol{I}_N - \boldsymbol{P}_{\boldsymbol{B}} - \frac{g}{1+g}\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{x}}{N} \tag{A.38}$$

$$m_N(\boldsymbol{x}) \triangleq \frac{\Gamma((N-l_N)/r)}{(N\pi\hat{\sigma}_N^2)^{(N-l_N)/r}|\boldsymbol{B}^H\boldsymbol{B}|^{1/r}} \tag{A.39}$$

where $\boldsymbol{P}_{\boldsymbol{B}}$ and $\boldsymbol{P}_{\boldsymbol{Z}}$ are the orthogonal projection matrices of $\boldsymbol{B}$ and $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$, respectively, and $\hat{\sigma}_k^2$ is asymptotically equal to the maximum likelihood (ML) estimate of the noise variance in the limit $\hat{\sigma}_{\mathrm{ML}}^2 = \lim_{g\to\infty}\hat{\sigma}_k^2$. The estimate $\hat{\sigma}_N^2$ is the estimated noise variance of the null model, and it is defined as $\hat{\sigma}_k^2$ for $\boldsymbol{P}_{\boldsymbol{Z}} = 0$. Finally, $m_N(\boldsymbol{x})$ is the unnormalised marginal likelihood of the null model. The linear parameters and the noise variance is now easily integrated out of the marginal likelihood. Doing this, we obtain that

$$p(\boldsymbol{x}|g,\boldsymbol{\phi}_k,\mathcal{M}_k) \propto \frac{m_N(\boldsymbol{x})}{(1+g)^{l_k/r}}\left(\frac{\hat{\sigma}_N^2}{\hat{\sigma}_k^2}\right)^{(N-l_N)/r} \tag{A.40}$$

$$= \frac{m_N(\boldsymbol{x})(1+g)^{(N-l_N-l_k)/r}}{(1+g[1-R_k^2(\boldsymbol{\phi}_k)])^{(N-l_N)/r}} \tag{A.41}$$

which we define as the unnormalised marginal likelihood $m_k(\boldsymbol{x}|g,\boldsymbol{\phi}_k)$ of model $\mathcal{M}_k$ given g and $\boldsymbol{\phi}_k$. Moreover,

$$R_k^2(\boldsymbol{\phi}_k) \triangleq \frac{\boldsymbol{x}^H\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{x}}{\boldsymbol{x}^H(\boldsymbol{I}_N - \boldsymbol{P}_{\boldsymbol{B}})\boldsymbol{x}} \tag{A.42}$$

resembles the coefficient of determination from classical linear regression analysis where it measures how well the data set fits the regression. Depending on whether we assume the hyperparameter g and non-linear parameters $\boldsymbol{\phi}_k$ to be fixed parameters or random variables, we also have to find values for them or integrate them out of the marginal likelihood in (A.41). We look into this issue in the next two sections.

4 Known System Matrix

In this section, we consider the case where there are either no non-linear parameters or they are known.

4.1 Fixed Choices of g

We first assume that g is a fixed quantity. From (A.6) and (A.40), the Bayes' factor is therefore

$$\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N|g,\boldsymbol{\phi}_k] = (1+g)^{-l_k/r}\left(\frac{\hat{\sigma}_N^2}{\hat{\sigma}_k^2}\right)^{(N-l_N)/r} . \tag{A.43}$$

With a uniform prior on the models, it follows from (A.7) that the Bayes' factor is proportional to the posterior distribution with pdf $p(\mathcal{M}_k|\boldsymbol{x},g,\boldsymbol{\phi}_k)$ on the models. The model with the highest posterior probability is the solution to

$$\hat{k} = \arg\max_{k\in\mathcal{K}} p(\mathcal{M}_k|\boldsymbol{x},g,\boldsymbol{\phi}_k) \tag{A.44}$$

$$= \arg\max_{k\in\mathcal{K}} \left[-(N-l_N)\ln\hat{\sigma}_k^2 - l_k\ln(1+g)\right] . \tag{A.45}$$

As alluded to in Sec. 3.1, the value of g is vital in model selection. From (A.41), we see that if $g\to\infty$, the Bayes' factor in (A.43) goes to zero. The null model is therefore always the most probable model, regardless of the information in the data (Bartlett's paradox). Another problem occurs if we assume that the least squares estimate $\boldsymbol{m}_k\to\infty$ or, equivalently, that $R_k^2(\boldsymbol{\phi}_k)\to 1$ so that the null model cannot be true. Although we would expect that the Bayes' factor would also go to infinity, it converges to the constant $(1+g)^{(N-l_n-l_k)/r}$, and this is called the information paradox [4, 35, 62]. For these two reasons, the value of g should depend on the data in some way. A local empirical Bayesian (EB) estimate is a data-dependent estimate of g, and it is the maximum of the marginal likelihood w.r.t. g [4]

$$g_k^{\mathrm{EB}} = \arg\max_{g\in\mathbb{R}^+} p(\boldsymbol{x}|g,\boldsymbol{\phi}_k,\mathcal{M}_k) \tag{A.46}$$

$$= \max\left(\frac{(N-l_N)R_k^2(\boldsymbol{\phi}_k)-l_k}{(1-R_k^2(\boldsymbol{\phi}_k))l_k},0\right) \tag{A.47}$$

where $\mathbb{R}^+$ is the set of non-negative real-valued numbers. This choice of g clearly resolves the information paradox. Inserting the EB estimate of g into (A.45) yields

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \left[-(N-l_N)\ln\frac{(N-l_N)\hat{\sigma}_{\mathrm{ML}}^2}{N-l_N-l_k} - l_k\ln(1+g_k^{\mathrm{EB}})\right] \tag{A.48}$$

whose form is similar to most of the information criteria. When the null model is the noise-only model so that $l_N=0$, these information criteria can be written as [20][5]

$$\hat{k} = \arg\max_{k\in\mathcal{K}} \left[-2N\ln\hat{\sigma}_{\mathrm{ML}}^2 - r\nu_k h(\nu_k,N)\right] \tag{A.49}$$

[5] The cost function must be divided by $2r$ when the information criteria are used for model averaging and comparison in the so-called multi-modal approach [21].

where ν_k is the number of real-valued independent parameters in the model, and $h(\nu_k, N)$ is a penalty coefficient. For $h(\nu_k, N) = \{2, \ln N\}$, we get the AIC and the BIC, respectively. Note that ν_k is not always the same as the number of unknown parameters [30]. Moreover, if the penalty coefficient is a linear function of ν_k or independent of it, ν_k may be interpreted as the number of independent parameters which are not in all candidate models. In nested models with white Gaussian noise such as the linear regression model considered in this section, this means that the noise variance parameter does not have to be counted as an independent parameter. Thus, selecting ν_k as either $\nu_k = 2l_k/r + 1$ or $\nu_k = 2l_k/r$ does not change, e.g., the AIC and the BIC. Rewriting (A.48) into the form of (A.49) and using (A.32) to show that $\eta_k^{\text{EB}} = l_k g_k^{\text{EB}}/N$ give the penalty coefficient

$$h(\nu_k, N) = \frac{2}{r\nu_k}[l_k \ln(1 + N\eta_k^{\text{EB}}/l_k) - N \ln(1 - l_k/N)] \tag{A.50}$$

which when inserted into (A.49) gives a model selection criterion we will refer to as the empirical BIC (EBIC).

Interpretation of the EBIC

To gain some insight into the behaviour of the EBIC, we here compare it to the AIC and the BIC in the context of a linear regression model with $N \gg l_k$ and $l_N = 0$. The number of independent parameters is therefore set to $\nu_k = 2l_k/r$ so that

$$h(\nu_k, N) = \ln(1 + N\eta_k^{\text{EB}}/l_k) - N \ln(1 - l_k/N)/l_k \tag{A.51}$$

$$\approx \ln(1 + N\eta_k^{\text{EB}}/l_k) + 1 \tag{A.52}$$

where the approximation follows from the assumption that $N \gg l_k$ so that $\ln(1-l_k/N) \approx -l_k/N$. From this approximation, several interesting observations can be made. When the SNR is large enough to justify that $N\eta_k^{\text{EB}} \gg l_k$, the EBIC is basically a corrected BIC which takes the estimated SNR of the data into account. The penalty coefficient grows with the estimated SNR and the chance of over-fitting thus becomes very low, even under high SNR conditions where the AIC, but also the BIC tend to overestimate the model order [63]. When the estimated SNR on the other hand becomes so low that $N\eta_k^{\text{EB}} \ll l_k$, the EBIC reduces to an AIC-like rule which has a constant penalty coefficient. In the extreme case of an estimated SNR of zero, the EBIC reduces to the so-called no-name rule [20]. Interestingly, empirical studies [39, 64] have shown that the AIC performs better than the BIC when the SNR in the data is low, and this is automatically captured by the EBIC. The EBIC therefore performs well across all SNR values as we demonstrate in Sec. 6.

4.2 Integration over g

Another way to resolve the information paradox is by treating g as a random variable and integrate it out of the marginal likelihood. For the prior distribution on g in (A.33)

and the unnormalised marginal likelihood in (A.41), we obtain a Bayes' factor given by

$$\begin{aligned}\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N|\boldsymbol{\phi}_k] &= \int_0^\infty \frac{m_k(\boldsymbol{x}|g,\boldsymbol{\phi}_k)}{m_N(\boldsymbol{x})}p(g|\delta)dg \\ &= \frac{\delta-r}{l_k+\delta-r}{}_2F_1\left(\frac{N-l_N}{r},1;\frac{l_k+\delta}{r};R_k^2(\boldsymbol{\phi}_k)\right) \end{aligned} \tag{A.53}$$

where ${}_2F_1$ is the Gaussian hypergeometric function [65, p. 314]. When N is large or $R_k^2(\boldsymbol{\phi}_k)$ is very close to one, numerical and computational problems with the evaluation of the Gaussian hypergeometric function may be encountered [66]. From a computational point of view, it may therefore not be advantageous to marginalise (A.53) w.r.t. g analytically. Instead, the Laplace approximation can be used as a simple alternative. Using the procedure outlined in Sec. 2.2 and the results in App. B, we get that

$$\mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N|\boldsymbol{\phi}_k] \approx \mathrm{BF}[\mathcal{M}_k;\mathcal{M}_N|\hat{g},\boldsymbol{\phi}_k]\frac{\hat{g}(\delta-r)}{r(1+\hat{g})^{\delta/r}}\sqrt{2\pi\gamma(\hat{g}|\boldsymbol{\phi}_k)} \tag{A.54}$$

where $\hat{g}=\exp(\hat{\tau})$ and $\gamma(\hat{g}|\boldsymbol{\phi}_k)$ can be found from (A.84) and (A.85), respectively, with $v=1$, $w=(N-l_N-l_k-\delta)/r$, and $u=(N-l_N)/r$. Since the marginal posterior distribution on g does not have a symmetric pdf and in order to avoid edge effects near $g=0$, the Laplace approximation was made for the parametrisation $\tau=\ln g$ [4]. This parametrisation suggests that the posterior distribution on g is approximately a log-normal distribution. The model with the highest posterior probability can be found by maximising (A.54) w.r.t. the model index and this yields the Laplace BIC (LP-BIC)

$$\hat{k}=\underset{k\in\mathcal{K}}{\arg\max}\left[-(N-l_N)\ln\hat{\sigma}_k^2-l_k\ln(1+\hat{g})-\delta\ln(1+\hat{g})+r\ln\hat{g}-r\gamma(\hat{g}|\boldsymbol{\phi}_k)/2\right] \tag{A.55}$$

Compared to the maximisation in (A.45), (A.55) differs in terms of the estimate of g and the last three terms. These terms accounts for the uncertainty in our point estimate of g.

5 Unkown Non-linear Parameters

In this section, the non-linear parameters $\boldsymbol{\phi}_k$ are also assumed unknown. First, $\boldsymbol{\phi}_k$ is considered an unknown but fixed quantity, and second, $\boldsymbol{\phi}_k$ is considered a random variable with the uniform prior in (A.34).

5.1 Estimating the non-linear Parameters

As in Sec. 4.1, we derive an EB estimator of the non-linear parameters given by

$$\begin{aligned}\hat{\boldsymbol{\phi}}_k^{\text{EB}} &= \arg\max_{\boldsymbol{\phi}_k \in \Phi_k} p(\boldsymbol{x}|g, \boldsymbol{\phi}_k, \mathcal{M}_k) = \arg\max_{\boldsymbol{\phi}_k \in \Phi_k} p(\boldsymbol{x}|\boldsymbol{\phi}_k, \mathcal{M}_k) \\ &= \arg\max_{\boldsymbol{\phi}_k \in \Phi_k} R_k^2(\boldsymbol{\phi}_k) = \arg\max_{\boldsymbol{\phi}_k \in \Phi_k} C_k(\boldsymbol{\phi}_k)\end{aligned} \tag{A.56}$$

where we have defined

$$C_k(\boldsymbol{\phi}_k) \triangleq \boldsymbol{x}^H \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{x} \ . \tag{A.57}$$

Note that $C_k(\boldsymbol{\phi}_k)$ does not depend on the hyperparameter g so the EB estimator $\hat{\boldsymbol{\phi}}_k^{\text{EB}}$ is independent of what we assume about g. Depending on the structure of $\boldsymbol{Z}_k(\boldsymbol{\phi}_k)$, it might be hard to perform the maximisation of $C_k(\boldsymbol{\phi}_k)$. In App. C, we have derived the first and second order differentials of an orthogonal projection matrix as these are useful in numerical optimisation algorithms for maximising $C_k(\boldsymbol{\phi}_k)$. For the non-linear regression model in (A.25), the EB estimator is identical to the ML estimator. Once an estimate of $\boldsymbol{\phi}_k$ has been found, the Bayes' factor can be computed by inserting this estimate into either (A.43), (A.53), or (A.54).

5.2 Integrating over the Non-linear Parameters

When we treat the ρ_k real-valued and non-linear parameters $\boldsymbol{\phi}_k$ as random variables, we must integrate them out of the marginal likelihood in (A.41). Since an analytical marginalisation is usually not possible, we here consider doing the joint integral

$$\text{BF}[\mathcal{M}_k; \mathcal{M}_N] = \int_{-\infty}^{\infty} \int_{\Phi_k} q(\boldsymbol{\phi}_k, \tau) d\boldsymbol{\phi}_k d\tau \tag{A.58}$$

using the Laplace approximation with the change of variables $\tau = \ln g$. The integrand is given by

$$q(\boldsymbol{\phi}_k, \tau) = \text{BF}[\mathcal{M}_k; \mathcal{M}_N | \exp(\tau), \boldsymbol{\phi}_k] p(\boldsymbol{\phi}_k|\mathcal{M}_k) p(\tau|\delta) \tag{A.59}$$

where

$$p(\tau|\delta) = \exp(\tau) p(g|\delta)\big|_{g=\exp(\tau)} \ . \tag{A.60}$$

For the uniform prior on $\boldsymbol{\phi}_k$ in (A.34), the mode of $q(\boldsymbol{\phi}_k, \tau)$ w.r.t. $\boldsymbol{\phi}_k$ is the EB estimate in (A.56). Evaluated at this mode, the Hessian matrix $\boldsymbol{H}(\boldsymbol{\phi}_k)$ is given by

$$\boldsymbol{H}(\hat{\boldsymbol{\phi}}_k^{\text{EB}}) = \frac{\exp(\tau)(N - l_N)}{rN[1 + \exp(\tau)]\hat{\sigma}_k^2} \boldsymbol{D} \tag{A.61}$$

where we have defined

$$\boldsymbol{D} \triangleq \frac{\partial^2 C_k(\boldsymbol{\phi}_k)}{\partial \boldsymbol{\phi}_k \partial \boldsymbol{\phi}_k^T}\bigg|_{\boldsymbol{\phi}_k = \hat{\boldsymbol{\phi}}_k^{\text{EB}}} \ . \tag{A.62}$$

Using the results in App. C, the (n,m)'th element of $\boldsymbol{D}$ can be written as

$$\begin{aligned}[\boldsymbol{D}]_{nm} = 2\mathrm{Re}\Big\{ & \boldsymbol{w}^H \big[\boldsymbol{\Lambda}_{nm} - \boldsymbol{T}_n \boldsymbol{S}_k^{-1} \boldsymbol{Z}_k^H(\hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}) \boldsymbol{T}_m \\ & - \boldsymbol{T}_m \boldsymbol{S}_k^{-1} \boldsymbol{Z}_k^H(\hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}) \boldsymbol{T}_n \big] \boldsymbol{m}_k + \boldsymbol{w}^H \boldsymbol{T}_n \boldsymbol{S}_k^{-1} \boldsymbol{T}_m^H \boldsymbol{w} \\ & - \boldsymbol{m}_k^H \boldsymbol{T}_n^H (\boldsymbol{I}_N - \boldsymbol{P}_{\boldsymbol{Z}}) \boldsymbol{T}_m \boldsymbol{m}_k \Big\} \end{aligned} \tag{A.63}$$

where we have defined

$$\boldsymbol{w} \triangleq \boldsymbol{x} - \boldsymbol{Z}_k(\hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}) \boldsymbol{m}_k \tag{A.64}$$

$$\boldsymbol{S}_k \triangleq \left[\boldsymbol{Z}_k^H(\hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}) \boldsymbol{Z}_k(\hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}) \right]^{-1} \tag{A.65}$$

$$\boldsymbol{T}_i \triangleq \left. \frac{\partial \boldsymbol{Z}_k(\boldsymbol{\phi}_k)}{\partial \phi_i} \right|_{\boldsymbol{\phi}_k = \hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}} \tag{A.66}$$

$$\boldsymbol{\Lambda}_{nm} \triangleq \left. \frac{\partial^2 \boldsymbol{Z}_k(\boldsymbol{\phi}_k)}{\partial \phi_n \partial \phi_m} \right|_{\boldsymbol{\phi}_k = \hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}} . \tag{A.67}$$

Since $\hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}$ does not depend on the value of τ, the mode and variance of $q(\boldsymbol{\phi}_k, \tau)$ w.r.t. τ is the same as in Sec. 4.2 and can be found in App. B with $v = 1$, $w = (N - l_N - l_k - \delta)/r$, and $u = (N - l_N)/r$. Thus, the Laplace approximation of the Bayes' factor in (A.58) is

$$\mathrm{BF}[\mathcal{M}_k; \mathcal{M}_N] \approx \mathrm{BF}[\mathcal{M}_k; \mathcal{M}_N | \hat{g}, \hat{\boldsymbol{\phi}}_k^{\mathrm{EB}}] \frac{\hat{g}(\delta - r)}{r(1+\hat{g})^{\delta/r}} \times W_k^{-1} (2\pi)^{(\rho_k+1)/2} \sqrt{\gamma(\hat{g}|\hat{\boldsymbol{\phi}}_k^{\mathrm{EB}})} |-\boldsymbol{H}(\hat{\boldsymbol{\phi}}_k^{\mathrm{EB}})|^{-1/2} . \tag{A.68}$$

When $q(\boldsymbol{\phi}_k, \tau)$ consists of multiple, significant, and well-separated peaks, the integral in (A.58) can be approximated by a Laplace approximation to each peak at their respective modes [52]. In this case, the Bayes' factor in (A.68) will be a sum over each of these peaks.

6 Simulations

We demonstrate the applicability of our model comparison algorithms by two simulation examples. In the first example, we compare the penalty coefficient of our proposed algorithms with the penalty coefficient of AIC, $\mathrm{AIC}_{\mathrm{c}}$, and BIC. In the second example, we consider model comparison in a periodic signal model which consists of a single non-linear parameter, the fundamental frequency. The simulation code can be found at `http://kom.aau.dk/~jkn/publications/publications.php`.

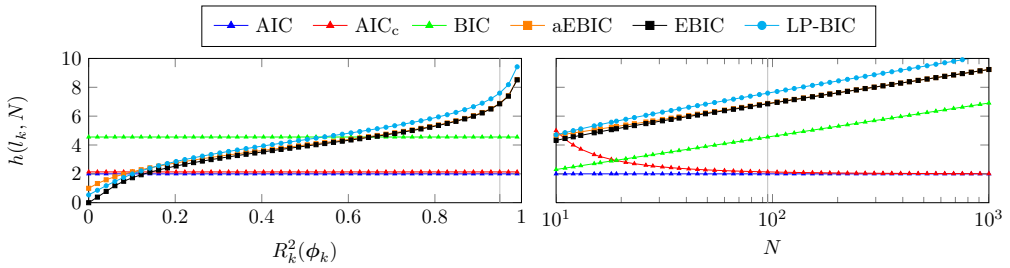

Fig. A.1: Interpretation of the various information criteria for $l_k = 5$. The plots show the penalty coefficient $h(l_k, N)$ as a function of the coefficient of determination $R_k^2(\phi_k)$ and the number of data points N. In the left plot, $N = 95$, and in the right, $R_k^2(\phi_k) = 0.95$ for the EBIC, the aEBIC, and the LP-BIC.

6.1 Penalty Coefficient

In Sec. 4.1, we considered the interpretation of the AIC and the BIC for a regression model with a known system matrix when the null model is the noise-only model and $N \gg l_k$. Here, we give some more insight by use of a simple simulation example in which the penalty coefficients of the AIC, the AIC$_c$, the BIC, the EBIC, the approximate EBIC (aEBIC), and the Laplace (LP-BIC) methods were found as a function of the coefficient of determination $R_k^2(\phi_k)$ and the number of data points N. The penalty coefficient of the aEBIC is given by (A.52). We fixed the number of linear parameters to $l_k = 5$, and Fig. A.1 shows the results.

In the left plot, the penalty coefficients $h(l_k, N)$ were computed as a function of the coefficient of determination for $N = 95$. Since the AIC, the AIC$_c$ and the BIC do not depend on the data, their penalty coefficients are constant. On the other hand, the penalty coefficients of the EBIC, the aEBIC, and the LP-BIC are data dependent and increase with the coefficient of determination.

In the right plot, the penalty coefficients $h(l_k, N)$ were computed as a function of the number of data points N for $R_k^2(\phi_k) = 0.95$. Note that BIC has the same trend as the EBIC, the aEBIC, and the LP-BIC although shifted. The vertical distance between BIC and EB or LP depends on the particular value of $R_k^2(\phi_k)$. In Fig. A.1, we set $R_k^2(\phi_k) = 0.95$, but if $R_k^2(\phi_k) \approx 0.648$ was selected instead, the EBIC and the BIC would coincide for large values of N.

6.2 Periodic Signal

We consider a complex periodic signal model given by

$$\mathcal{M}_k : \quad x(n) = \sum_{i=1}^{L} \alpha_i \exp(j\omega i n) \mathbb{I}_{\mathcal{L}_k}(i) + e(n) \tag{A.69}$$

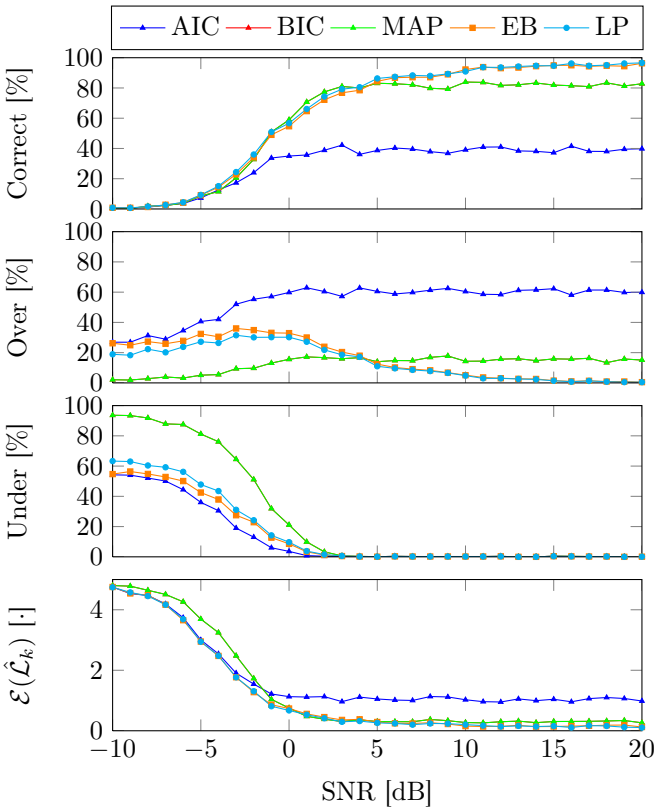

Fig. A.2: Percentage of correctly detected models, overestimated models, underestimated models, and the MSE of the estimated model versus the SNR. The model was a periodic signal model with a maximum of $L = 10$ harmonics resulting in $2^{10} - 1 = 1023$ different models. The curves corresponding to BIC and MAP are almost coinciding.

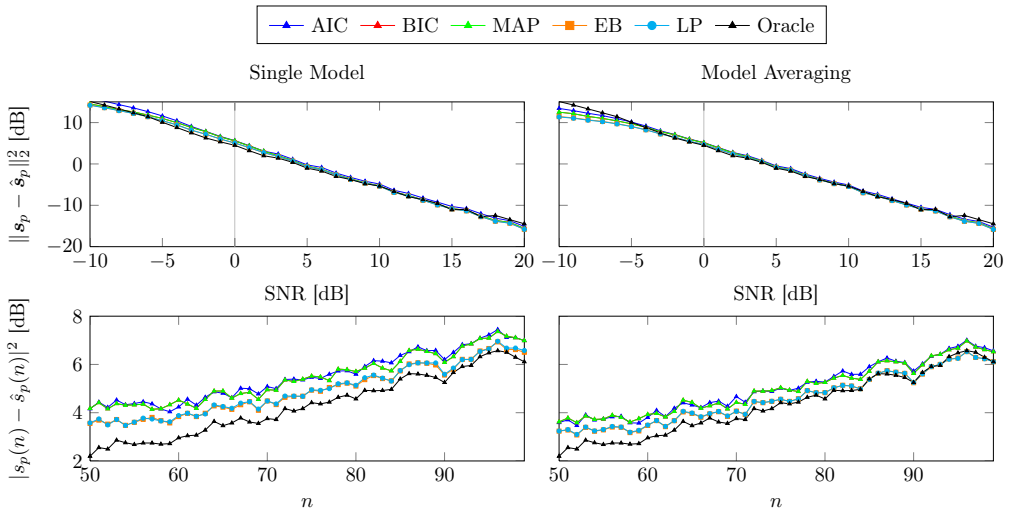

Fig. A.3: Prediction performance versus the SNR (top row) and versus the prediction step at an SNR of 0 dB (bottom row) for a periodic signal model. In the plots in the left column, only the model with the highest posterior probability is used. In the plots in the right column, all models are used for making the prediction. The curves corresponding to BIC and MAP are almost coinciding.

for $n = 0, 1, \ldots, N-1$ where $\mathbb{I}_{\mathcal{L}_k}(i)$ indicates whether the i'th harmonic component is included in the model $\mathcal{M}_k$ or not. This model is a special case of the model in (A.25) with the null model being the noise-only model, $\phi_k = \omega$, and $\boldsymbol{\alpha}_k$ being the complex amplitudes. Since no closed-form solution exists for the posterior distribution on the models for the periodic signal model, we consider the two approximations suggested in Sec. 5. We refer to the approximation based on (A.43) as the EB method and the approximation based on (A.68) as the Laplace (LP) method. The methods are compared to the AIC, the BIC, and the asymptotic MAP rule by Djuric with the latter having the penalty coefficient [9]

$$h(l_k, N) = \ln N + \frac{3}{2l_k} \ln N \ . \tag{A.70}$$

For the periodic signal model, the Hessian matrix in (A.61) is a scalar which can be approximated by [67]

$$H(\hat{\omega}_k^{\mathrm{EB}}) = -\frac{\hat{g} N (N^2-1) \sum_{i=1}^{L} |[\boldsymbol{m}_k]_i|^2 i^2 \mathbb{I}_{\mathcal{L}_k}(i)}{6(1+\hat{g})\hat{\sigma}_k^2} \ . \tag{A.71}$$

In the simulations, we set the maximum number of harmonic components to $L = 10$ and considered $K = 2^L - 1 = 1023$ models. Zero prior probability was assigned to

the noise-only model as the model comparison performance was evaluated against the SNR. Moreover, this permits the use of the improper prior $p(g|\delta = r = 1)$ since g is now a common parameter in all models. For each SNR from -10 dB to 20 dB in steps of 1 dB, we ran 1000 Monte Carlo runs. As recommended in [21], a data vector $\boldsymbol{x}$ consisting of $N = 50$ samples was generated in each run by first randomly selecting a model from a uniform prior on the models. For this model, we then randomly selected the fundamental frequency and the phases of the complex amplitudes from a uniform distribution on the interval $[(2\pi)^{-1}, 2\pi/\max(\mathcal{L}_k)]$ and $[0, 2\pi]$, respectively. The amplitudes of the harmonics in the selected model were all set to one. Finally a complex noise vector was generated and normalised so that the data had the desired SNR. Besides generating a data vector, we also generated a vector $\boldsymbol{x}_p$ of unobserved data for $n = N, N+1, \ldots 2N-1$.

In Fig. A.2, the percentage of correctly detected models, overestimated models, underestimated models, and the mean-squared-error (MSE) of the estimated model versus the SNR is shown. The MSE is defined as

$$\mathcal{E}(\hat{\mathcal{L}}_k) = \sum_{i=1}^{L} (\mathbb{I}_{\mathcal{L}_k}(i) - \mathbb{I}_{\hat{\mathcal{L}}_k}(i))^2 \tag{A.72}$$

where $\hat{\mathcal{L}}_k$ is the set containing the harmonic numbers of the most likely model. For an SNR above 0 dB, the EB and LP methods have almost identical performance with the LP method performing slightly better. The BIC and the MAP method are visually indistinguishable and perform generally worse than the EB and LP methods. AIC performs much worse than all other methods. For SNRs below 0 dB, the BIC and the MAP method have a strong tendency to underestimate the model whereas the underestimation tendency is only weak for the other methods. In terms of the MSE, the BIC and the MAP method performs worse than the other methods for low SNRs. However, it should be noted that the percentage of correctly detected models is not necessarily the best way of benchmarking model selection methods. As exemplified in [21], the true model does not always give the best prediction performance, and it may therefore be advantageous to either over- or underestimate the model order.

We have therefore also investigated the prediction performance, and the results are shown in Fig. A.3. In the plots in the left column, only the single model with the largest posterior probability was used for making the predictions of the predictable part $\boldsymbol{s}_p$ whereas all models were used as in (A.8) in the plots in the right column. The prediction based on a single model and all models was the mean of $p(\boldsymbol{x}_p|\boldsymbol{x}, \mathcal{M}_k)$ and $p(\boldsymbol{x}_p|\boldsymbol{x})$, respectively, where the latter depends on the former as in (A.8) with

$$p(\boldsymbol{x}_p|\boldsymbol{x}, \mathcal{M}_k) = \int_{\Theta_k} p(\boldsymbol{x}_p|\boldsymbol{\theta}_k, \mathcal{M}_k) p(\boldsymbol{\theta}_k|\boldsymbol{x}, \mathcal{M}_k) d\boldsymbol{\theta}_k \ . \tag{A.73}$$

We have evaluated this integral by treating the fundamental frequency as a fixed and unknown variable and by using the normal approximation on $\tau = \ln g$. In the top row,

the MSE of the total prediction error versus the SNR is shown, and in the bottom row, the MSE of the prediction error for each prediction step at an SNR of 0 dB is shown. In the four plots, the oracle knows the true model but not the model parameters. From the four figures, we see again that EB and LP outperform the other methods with AIC being the overall worst. For low SNRs, we also see that the MSE of the prediction errors is significantly lower when model averaging is used. Moreover, we see that the performance is also better than the oracle performance and this demonstrates, as discussed above, that the true model does not always give the best prediction performance. For high SNRs, only AIC performs slightly worse than the other methods which performs almost as well as the oracle. Moreover, there is basically no difference between the single and multi-model predictions since a single model receives all posterior probability.

7 Conclusion

Model comparison and selection is a difficult and important problem and a lot of methods have therefore been proposed. In this paper, we first gave an overview over how model comparison is performed for any model in a Bayesian framework. We also discussed the two major issues of doing the model comparison in a Bayesian framework, namely the elicitation of prior distributions and the evaluation of the marginal likelihood. Specifically, we reviewed the conditions for using improper prior distributions, and we briefly discussed approximate numerical and analytical algorithms for evaluating the marginal likelihood. In the second part of the paper, we analysed a general regression model in a Bayesian framework. The model consisted of both linear and non-linear parameters, and we used and motivated a prior of the same form as the Zellner's g-prior for this model. Many of the information criteria can be interpreted in a new light using this model with known non-linear parameters. These interpretations also gave insight into why the AIC and the AIC_c often overestimate the model complexity for a high SNR, and why the BIC underestimate the model complexity for a low SNR. For unknown non-linear parameters, we proposed an approximate way of integrating them out of the marginal likelihood using the Laplace approximation, and we demonstrated through a simple simulation example that our proposed model comparison and selection algorithms outperform other algorithms such as the AIC, the BIC and the asymptotic MAP rule both in terms of detecting the true model and in making predictions.

A Fisher Information Matrix for the Observation Model

Let $\boldsymbol{\gamma}$ denote a mixed parameter vector of complex-valued and real-valued parameters. Using the procedure in [68, App. 15C], it can be shown that the (n, m)'th element of the Fisher information matrix (FIM) for the normal distribution $\mathcal{N}_r(\boldsymbol{x}; \boldsymbol{\mu}(\boldsymbol{\gamma}); \boldsymbol{\Sigma}(\boldsymbol{\gamma}))$ is

given by

$$[\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})]_{nm} = \frac{1}{r}\left(\frac{\partial \boldsymbol{\mu}^*(\boldsymbol{\gamma})}{\partial \gamma_n^*}\right)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})\left(\frac{\partial \boldsymbol{\mu}^*(\boldsymbol{\gamma})}{\partial \gamma_m^*}\right)^* + \frac{1}{r}\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\gamma})}{\partial \gamma_n^*}\right)^H \boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})\left(\frac{\partial \boldsymbol{\mu}(\boldsymbol{\gamma})}{\partial \gamma_m^*}\right) + \frac{1}{r}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\gamma})}{\partial \gamma_n^*}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\gamma})\left(\frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\gamma})}{\partial \gamma_m^*}\right)^H\right). \tag{A.74}$$

For the observation model in (A.29), the parameter vector is given by $\boldsymbol{\gamma} = \begin{bmatrix}\boldsymbol{\psi}^T & \boldsymbol{\alpha}_k^T & \boldsymbol{\phi}_k^T & \sigma^2\end{bmatrix}^T$, and the mean vector and covariance matrix are given by

$$\boldsymbol{\mu}(\boldsymbol{\gamma}) = \boldsymbol{B}\boldsymbol{\psi} + \boldsymbol{Z}_k(\boldsymbol{\phi}_k)\boldsymbol{\alpha}_k \tag{A.75}$$

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \sigma^2 \boldsymbol{I}_N \ . \tag{A.76}$$

Computing the derivatives in (A.74) for the observation model in (A.29) yields the FIM given by

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma}) = \frac{1}{\sigma^2}\begin{bmatrix}\boldsymbol{B}^H\boldsymbol{B} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mathcal{I}}(\boldsymbol{\alpha}_k, \boldsymbol{\phi}_k) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \frac{N}{r\sigma^2}\end{bmatrix} \tag{A.77}$$

where

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\alpha}_k, \boldsymbol{\phi}_k) = \begin{bmatrix}\boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k) & \boldsymbol{Z}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Q}_k(\boldsymbol{\phi}_k) \\ \boldsymbol{Q}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Z}_k(\boldsymbol{\phi}_k) & \frac{2}{r}\operatorname{Re}\left(\boldsymbol{Q}_k^H(\boldsymbol{\phi}_k)\boldsymbol{Q}_k(\boldsymbol{\phi}_k)\right)\end{bmatrix} \tag{A.78}$$

$$\boldsymbol{Q}_k(\boldsymbol{\phi}_k) \triangleq \frac{\partial(\boldsymbol{Z}_k(\boldsymbol{\phi}_k)\boldsymbol{\alpha}_k)}{\partial \boldsymbol{\phi}_k} \ . \tag{A.79}$$

Note that $\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})$ is block diagonal which follows from the assumption that $\boldsymbol{B}^H\boldsymbol{Z}_k(\boldsymbol{\phi}_k) = \boldsymbol{0}$.

B Laplace Approximation with the Hyper-g Prior

For the hyper-g prior in (A.33), the integral in (A.53) with the change of variables to $\tau = \ln g$ can be written in the form

$$\int_0^\infty g^{v-1}(1+g)^w[1+g(1-R_k^2(\boldsymbol{\phi}_k))]^{-u}dg = \int_{-\infty}^\infty \exp(v\tau)(1+\exp(\tau))^w[1+\exp(\tau)(1-R_k^2(\boldsymbol{\phi}_k))]^{-u}d\tau \ . \tag{A.80}$$

Taking the derivative of the logarithm of the integrand and equating to zero leads to the quadratic equation

$$0 = \alpha_\tau \exp(2\tau) + \beta_\tau \exp(\tau) + v \tag{A.81}$$

where we have defined

$$\alpha_\tau \triangleq (1 - R_k^2(\boldsymbol{\phi}_k))(v + w - u) \tag{A.82}$$

$$\beta_\tau \triangleq (u - v)R_k^2(\boldsymbol{\phi}_k) + 2v + w - u \tag{A.83}$$

For $u - w > v$, the only positive solution to this quadratic equation is

$$\hat{\tau} = \ln\left(\frac{\beta_\tau + \sqrt{\beta_\tau^2 - 4\alpha_\tau v}}{-2\alpha_\tau}\right) \tag{A.84}$$

which is the mode of the normal approximation to the integrand. The corresponding variance at this mode with $\hat{g} = \exp(\hat{\tau})$ is

$$\gamma(\hat{g}|\boldsymbol{\phi}_k) = \left[\frac{\hat{g}u(1 - R_k^2(\boldsymbol{\phi}_k))}{(1+\hat{g})^2} - \frac{\hat{g}w}{[1 + \hat{g}(1 - R_k^2(\boldsymbol{\phi}_k))]^2}\right]^{-1} . \tag{A.85}$$

C Differentials of a Projection Matrix

Let $\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{G}^H\boldsymbol{G})^{-1}\boldsymbol{G}^H$ denote an orthogonal projection matrix, and let $\boldsymbol{S} = \boldsymbol{G}^H\boldsymbol{G}$ denote an inner matrix product. The differential of $\boldsymbol{S}$ is then given by

$$d\boldsymbol{S} = (d\boldsymbol{G})^H\boldsymbol{G} + \boldsymbol{G}^H(d\boldsymbol{G}) . \tag{A.86}$$

This result can be used to show that

$$d\boldsymbol{S}^{-1} = -\boldsymbol{S}^{-1}[(d\boldsymbol{G})^H\boldsymbol{G} + \boldsymbol{G}^H(d\boldsymbol{G})]\boldsymbol{S}^{-1} , \tag{A.87}$$

and that

$$d\boldsymbol{P} = \boldsymbol{P}^\perp(d\boldsymbol{G})\boldsymbol{S}^{-1}\boldsymbol{G}^H + \boldsymbol{G}\boldsymbol{S}^{-1}(d\boldsymbol{G})^H\boldsymbol{P}^\perp \tag{A.88}$$

where $\boldsymbol{P}^\perp = \boldsymbol{I} - \boldsymbol{P}$ is the complementary projection of $\boldsymbol{P}$. Let δ denote another differential operator. From the above results, we obtain after some algebra that

$$\begin{aligned}
\delta(d\boldsymbol{P}) &= \boldsymbol{P}^\perp(\delta(d\boldsymbol{G}))\boldsymbol{S}^{-1}\boldsymbol{G}^H + \boldsymbol{G}\boldsymbol{S}^{-1}(\delta(d\boldsymbol{G}))^H\boldsymbol{P}^\perp \\
&+ \boldsymbol{P}^\perp\left[(d\boldsymbol{G})\boldsymbol{S}^{-1}(\delta\boldsymbol{G})^H + (\delta\boldsymbol{G})\boldsymbol{S}^{-1}(d\boldsymbol{G})^H\right]\boldsymbol{P}^\perp \\
&- \boldsymbol{P}^\perp\left[(\delta\boldsymbol{G})\boldsymbol{S}^{-1}\boldsymbol{G}^H(d\boldsymbol{G}) + (d\boldsymbol{G})\boldsymbol{S}^{-1}\boldsymbol{G}^H(\delta\boldsymbol{G})\right]\boldsymbol{S}^{-1}\boldsymbol{G}^H \\
&- \boldsymbol{G}\boldsymbol{S}^{-1}\left[(\delta\boldsymbol{G})^H\boldsymbol{G}\boldsymbol{S}^{-1}(d\boldsymbol{G})^H + (d\boldsymbol{G})^H\boldsymbol{G}\boldsymbol{S}^{-1}(\delta\boldsymbol{G})^H\right]\boldsymbol{P}^\perp \\
&+ \boldsymbol{G}\boldsymbol{S}^{-1}\left[(d\boldsymbol{G})^H\boldsymbol{P}^\perp(\delta\boldsymbol{G}) + (\delta\boldsymbol{G})^H\boldsymbol{P}^\perp(d\boldsymbol{G})\right]\boldsymbol{S}^{-1}\boldsymbol{G}^H .
\end{aligned} \tag{A.89}$$

References

[1] G. E. P. Box and N. R. Draper, *Empirical model-building and response surface*. John Wiley & Sons, Inc., Jan. 1987.

[2] E. I. George and D. P. Foster, "Calibration and empirical Bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747, Dec. 2000.

[3] M. Clyde and E. I. George, "Model uncertainty," *Statist. Sci.*, vol. 19, no. 1, pp. 81–94, Feb. 2004.

[4] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, "Mixtures of g priors for Bayesian variable selection," *J. Amer. Statistical Assoc.*, vol. 103, pp. 410–423, Mar. 2008.

[5] L. Kavalieris and E. J. Hannan, "Determining the number of terms in a trigonometric regression," *J. of Time Series Analysis*, vol. 15, no. 6, pp. 613–625, Nov. 1994.

[6] B. G. Quinn, "Estimating the number of terms in a sinusoidal regression," *J. of Time Series Analysis*, vol. 10, no. 1, pp. 71–75, Jan. 1989.

[7] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2667–2676, 1999.

[8] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2498–2517, Apr. 2006.

[9] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, B. H. Juang, Ed. Morgan & Claypool, 2009.

[10] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Royal Stat. Soc., Series B*, vol. 41, no. 2, pp. 190–195, 1979.

[11] G. Liang, D. M. Wilkes, and J. A. Cadzow, "Arma model order estimation based on the eigenvalues of the covariance matrix," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3003–3009, Oct. 1993.

[12] B. Choi, *ARMA model identification*. Springer-Verlag, Jun. 1992.

[13] S. Koreisha and G. Yoshimoto, "A comparison among identification procedures for autoregressive moving average models," *International Statistical Review*, vol. 59, no. 1, pp. 37–57, Apr. 1991.

[14] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Reversible jump Markov chain Monte Carlo strategies for Bayesian model selection in autoregressive processes," *J. of Time Series Analysis*, vol. 25, no. 6, pp. 785–809, Nov. 2004.

[15] T. Cassar, K. P. Camilleri, and S. G. Fabri, "Order estimation of multivariate ARMA models," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 3, pp. 494–503, Jun. 2010.

[16] Z. Liang, R. Jaszczak, and R. Coleman, "Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing," *IEEE Trans. Nucl. Sci.*, vol. 39, no. 4, pp. 1126–1133, Aug. 1992.

[17] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Adv. in Neural Inf. Process. Syst.*, 2000, pp. 554–560.

[18] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[19] C. R. Rao and Y. Wu, "On model selection," *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, vol. 38, pp. 1–57, 2001.

[20] P. Stoica and Y. Selén, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[21] P. Stoica, Y. Selén, and J. Li, "Multi-model approach to model selection," *Digital Signal Process.*, vol. 14, no. 5, pp. 399–412, Sep. 2004.

[22] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Prentice Hall, May 2005.

[23] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[24] C. M. Hurvich and C.-L.Tsai, "A corrected Akaike information criterion for vector autoregressive model selection," *J. of Time Series Analysis*, vol. 14, no. 3, pp. 271–279, May 1993.

[25] S. Konishi and G. Kitagawa, "Generalised information criteria in model selection," *Biometrika*, vol. 83, no. 4, pp. 875–890, Dec. 1996.

[26] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[27] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978.

[28] ——, "Estimation of structure by minimum description length," *Circuits, Systems, and Signal Process.*, vol. 1, no. 3, pp. 395–406, 1982.

[29] ——, "A predictive least-squares principle," *IMA J. Math. Control Inf.*, vol. 3, no. 2–3, pp. 211–222, 1986.

[30] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[31] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 450–458, Feb. 2006.

[32] J.-M. Papy, L. De Lathauwer, and S. Van Huffel, "A shift invariance-based order-selection technique for exponential data modelling," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 473–476, Jul. 2007.

[33] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. on Advances in Signal Process.*, 2009.

[34] J. O. Berger and L. R. Pericchi, "The intrinsic Bayes factor for model selection and prediction," *J. Amer. Statistical Assoc.*, vol. 91, no. 433, pp. 109–122, Mar. 1996.

[35] ——, "Objective Bayesian methods for model selection: Introduction and comparison," *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, vol. 38, pp. 135–207, 2001.

[36] L. Wasserman, "Bayesian model selection and model averaging," *J. of Mathematical Psychology*, vol. 44, no. 1, pp. 92–107, Mar. 2000.

[37] A. F. Dentell, "Objective bayes criteria for variable selection," Ph.D. dissertation, Universitat de Valencia, 2011.

[38] P. M. Djuric, "A model selection rule for sinusoids in white Gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, Jul. 1996.

[39] ——, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[40] A. Zellner, "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian Inference and Decision Techniques*. Elsevier, 1986.

[41] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Jun. 2002.

[42] R. Strachan and H. K. v. Dijk, "Improper priors with well defined Bayes' factors," Department of Economics, University of Leicester, Discussion Papers in Economics 05/4, 2005.

[43] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, May 2001.

[44] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer-Verlag New York, Inc., Jul. 2004.

[45] C. Andrieu, A. Doucet, and C. P. Robert, "Computational advances for and from Bayesian analysis," *Statist. Sci.*, vol. 19, no. 1, pp. 118–127, Feb. 2004.

[46] S. Chib, "Marginal likelihood from the Gibbs output," *J. Amer. Statistical Assoc.*, vol. 90, no. 432, pp. 1313–1321, Dec. 1995.

[47] S. Chib and I. Jeliazkov, "Marginal likelihood from the Metropolis-Hastings output," *J. Amer. Statistical Assoc.*, vol. 96, no. 453, pp. 270–281, Mar. 2001.

[48] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.

[49] C. Han and B. P. Carlin, "Markov chain Monte Carlo methods for computing Bayes factors: A comparative review," *J. Amer. Statistical Assoc.*, vol. 96, no. 455, pp. 1122–1132, Sep. 2001.

[50] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, Aug. 2006.

[51] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal," *J. Amer. Statistical Assoc.*, vol. 81, no. 393, pp. 82–86, Mar. 1986.

[52] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Jul. 2003.

[53] A. Zellner and A. Siow, "Posterior odds ratios for selected regression hypotheses," *Trabajos de Estadística y de Investigación Operativa*, vol. 31, pp. 585–603, 1980.

[54] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, no. 3, pp. 227–241, 1968.

[55] ——, *Probability Theory : The Logic of Science*, G. L. Bretthorst, Ed. Cambridge University Press, Apr. 2003.

[56] G. L. Bretthorst, "The near-irrelevance of sampling frequency distributions," in *Max. Entropy and Bayesian Methods*, 1999, pp. 21–46.

[57] E. T. Jaynes, "Bayesian spectrum and chirp analysis," in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G. J. Erickson, Eds. D. Reidel, Dordrecht-Holland, 1987, pp. 1–37.

[58] D. S. Bové and L. Held, "Hyper-g priors for generalized linear models," *Bayesian Analysis*, vol. 6, no. 3, pp. 387–410, 2011.

[59] C. Fernández, E. Ley, and M. F. J. Steel, "Benchmark priors for Bayesian model averaging," *J. Econometrics*, vol. 100, no. 2, pp. 381–427, Feb. 2001.

[60] W. Cui and G. E. I., "Empirical Bayes vs. fully Bayes variable selection," *J. Stat. Planning and Inference*, vol. 138, no. 4, pp. 888–900, Apr. 2008.

[61] R. Guo and P. L. Speckman, "Bayes factor consistency in linear models," in *The 2009 International Workshop on Objective Bayes Methodology*, Jun. 2009.

[62] A. Zellner, "Comments on 'Mixtures of g-priors for Bayesian Variable Selection' by F. Liang, R. Paulo, G. Molina, M.A. Clyde and J.O. Berger," Jul. 2008.

[63] Q. Ding and S. M. Kay, "Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1959–1969, May 2011.

[64] Q. T. Zhang and K. M. Wong, "Information theoretic criteria for the determination of the number of signals in spatially correlated noise," *IEEE Trans. Signal Process.*, vol. 41, no. 4, pp. 1652–1663, Apr. 1993.

[65] I. S. Gradshteĭn, I. M. Ryzhik, and A. Jeffrey, *Table of Integrals, Series, and Products*. Academic Press, 2000.

[66] R. W. Butler and A. T. A. Wood, "Laplace approximations for hypergeometric functions with matrix argument," *Ann. Stat.*, vol. 30, no. 4, pp. 1155–1177, Aug. 2002.

[67] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "An approximate Bayesian fundamental frequency estimator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012.

[68] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall PTR, Mar. 1993.

Paper B

Default Bayesian Estimation of the Fundamental Frequency

Jesper Kjær Nielsen, Mads Græsbøll Christensen, and Søren Holdt Jensen

The paper has been submitted to the
IEEE Transactions on Audio, Speech, and Language Processing,

In peer-review
The layout has been revised.

Abstract

Joint fundamental frequency and model order estimation is an important problem in several applications. In this paper, a default estimation algorithm based on a minimum of prior information is presented. The algorithm is developed in a Bayesian framework, and it can be applied to both real- and complex-valued discrete-time signals which may have missing samples or may have been sampled at a non-uniform sampling frequency. For the elicitation of a prior distribution, a prior of the same form as the Zellner's g-prior is demonstrated to be a good approximation to a default prior distribution, and several approximations of the posterior distributions on the fundamental frequency and the model order are derived. Moreover, one of the state-of-the-art joint fundamental frequency and model order estimators is shown to be a special case of one of these approximations. The performance of the approximations are evaluated in a small-scale simulation study on both synthetic and real world signals. The simulations indicate that the proposed algorithm yields more accurate results than previous algorithms.

1 Introduction

An important and basic problem in time-series analysis is the estimation of the fundamental frequency and the number of harmonic components of a periodic signal. The problem is encountered in a wide range of science and engineering applications such as music processing [1, 2], speech processing [3, 4], sonar [5], electrocardiography (ECG) [6], and seismology [7]. In particular for musical applications, fundamental frequency estimation has been subject to extensive research for several decades [2]. This is primarily due to that a musical note is composed of the sum of a fundamental partial and a number of overtone partials. For harmonic instruments, these overtone partials are called harmonics since their frequencies $\{\omega_i\}_{i=2}^{l}$ are approximately related to the fundamental frequency ω of the fundamental partial by $\omega_i \approx i\omega$ for $i = 2, \cdots, l$ [1, 8]. Since the fundamental frequency is such an important physical attribute to musical applications, the more elegant term pitch is often used instead [2]. Therefore, the problem considered in this paper is often referred to as (single-)pitch estimation in the context of musical applications.

The problem of estimating the fundamental frequency is typically defined in the following way. A data set $\{x(t_n)\}_{n=0}^{N-1}$ originating from a discrete-time signal is observed and modelled as

$$x(t_n) = s(t_n) + e(t_n) \ , \quad n = 0, 1, \cdots, N-1 \tag{B.1}$$

where $\{t_n\}_{n=0}^{N-1}$, $\{s(t_n)\}_{n=0}^{N-1}$, and $\{e(t_n)\}_{n=0}^{N-1}$ are the sampling times, the predictable part of the signal, and the non-predictable part of the signal, respectively. Usually, the sampling period T is assumed to be constant so that $t_n = nT$ for $t_0 = 0$. However, in

order to allow for a non-uniform sampling scheme or missing samples, this assumption is not made here. The predictable part consists of l harmonic components and is at time t_n given by

$$s(t_n) = \begin{cases} \sum_{i=1}^{l} \alpha_i \exp(ji\omega t_n), & x(t_n) \in \mathbb{C} \\ \sum_{i=1}^{l} a_i \cos(i\omega t_n) + b_i \sin(i\omega t_n), & x(t_n) \in \mathbb{R} \end{cases} \tag{B.2}$$

where $\mathbb{C}$ and $\mathbb{R}$ denote the set of complex and real numbers, respectively, and $j = \sqrt{-1}$ is the imaginary unit. For the i'th harmonic component, the complex amplitude α_i, the in-phase component a_i, the quadrature components b_i, the amplitude A_i, and the phase ϕ_i are related by

$$\alpha_i = a_i + jb_i = A_i \exp(j\phi_i) \ . \tag{B.3}$$

Note that a real-valued signal of the form in (B.2) can be cast into the form of a complex-valued signal in (B.2) by computing its down-sampled analytic signal [9]. Provided that the frequencies of the first and last harmonics are not too close to zero and the Nyquist frequency (relative to N), respectively, the solution to the estimation problem using the down-sampled analytic signal yields nearly the same result as for the real-valued signal [2, 10]. In this paper, the focus is on the complex-valued signal model since it leads to simpler notation and faster algorithms [2, 11]. However, the results for the real-valued signal model is also given since the transformations from real-valued to analytic signals and vice versa under non-uniform sampling or with missing samples are more time consuming.

Numerous fundamental frequency estimation algorithms have been suggested in the literature. The simplest algorithms are the non-parametric methods based on, for example, the auto-correlation function [12, 13] or the cepstrum [14] (See [15, 16] for other non-parametric methods). The more advanced algorithms are based on a signal model of the observed signal and are therefore referred to as parametric methods. These are typically maximum likelihood-based (ML) methods [17, 18], subspace-based methods [11, 19], filtering methods [20, 21], or Bayesian methods [8, 22, 23]. We refer the interested reader to [2] for a review of many of the non-Bayesian methods. Only a few of the suggested methods assume that the number of harmonics is unknown. In order to perform model selection, these methods typically add an order dependent penalty term to the log-likelihood function [24–26], use the eigenvalues [27] or eigenvectors [28, 29], or compare the angle between subspaces [30]. A good overview over these and other methods can be found in [2]. In contrast to model comparison in which a probability for each model is computed, these methods are typically only designed for detecting the most likely model. On the other hand, model comparison enables us to account for model uncertainty in the estimation of unknown model parameters and the predic-

tion of missing data points by using all models instead of just the most likely one. As demonstrated in, e.g., [31], model averaging increases the prediction performance.

In this paper, inference about the fundamental frequency and the number of harmonics are made in a Bayesian framework. The Bayesian framework is used for model comparison since it leads to consistent estimates under very mild conditions, naturally selects the simplest model which explains the data reasonably well (the principle of Occam's razor [32]), takes model uncertainty into account for estimation and prediction, and enables a more intuitive interpretation of the results [33, 34]. In a Bayesian framework, prior distributions on the unknown quantities must be elicited and their hyperparameters must be selected. In general, this is not a trivial problem since the prior information is usually not in the form of probability distributions, and prior information must therefore be turned into one or several probability distributions. For model comparison, this prior elicitation is very important since improper or vague priors may lead to indeterminate or bad answers [34]. Another difficulty of the Bayesian methods is that closed-form analytical solutions usually do not exist. Various numerical algorithms such as Markov chain Monte Carlo sampling [35] can overcome this limitation, but the computational load of running these algorithms is typically very high.

The primary aim of this paper is to develop a default estimation scheme for estimating the fundamental frequency and the number of harmonics. Note that even though the number of harmonics might not be of interest by itself, it is still vital to estimate it in order to avoid problems with for example pitch halving [11]. By the word *default*, we mean that an almost user-parameter free algorithm is developed which automatically follows from a minimum of prior information and a few minor approximations. The approximations are made so that closed-form expressions are obtained which have a computational load comparable to the methods suggested in [2]. Moreover, we show that a special case of the proposed approach is identical to the algorithm proposed in [2, Sec. 2.6]. Finally, we demonstrate through simulation examples that the proposed method is superior to the state-of-the-art ML-based and subspace-based methods. Note that we are here not concerned with the development of a full pitch detection and tracking system for speech or music applications such as YIN [12], RAPT [36], or NDF [37]. However, we believe that our estimator may be a useful component in such systems as well as in similar systems for other application domains.

The paper is organised as follows. The primary aim and the inference method are presented in Sec. 2. In Sec. 3, the default observation model and prior distributions are developed, and it is shown that these can be approximated by a prior distribution with the same form as the Zellner's g-prior. The observation model and the prior distributions are turned into posterior distributions on the fundamental frequency and the model order in Sec. 4. In Sec. 5, various approximations of varying accuracy and computational load are developed, and in Sec. 6 it is demonstrated that a state-of-the-art ML-based algorithm is a special case of one of these approximations. In Sec. 7, the approximations are evaluated on a synthetic signal, and the applicability of the

algorithm is demonstrated for the spectral analysis of a speech signal. Finally, Sec. 8 concludes this paper.

2 Problem Formulation and Background

The primary aim is to make inference about the fundamental frequency ω and the model order l given the prior information I and the N data points collected in the vector $\boldsymbol{x}$. That is, we wish to find the posterior densities

$$p(\omega, l|\boldsymbol{x}, I) = p(\omega|\boldsymbol{x}, l, I)p(l|\boldsymbol{x}, I) \tag{B.4}$$

and some of their statistics such as the mode, the mean, and the variance. In (B.4) and the rest of the paper, the generic notation $p(\cdot)$ is used to denote both a probability density function (pdf) over a continuous parameter and a probability mass function (pmf) over a discrete parameter. The model order l labels a unique model $\mathcal{M}_l$ with model parameters $\boldsymbol{\theta}_l \in \Theta_l$. For the problem at hand, ω is one of these parameters, and the remaining model parameters such as the noise parmeters and the complex amplitudes are nuisance parameters. The observation model $p(\boldsymbol{x}|\boldsymbol{\theta}_l, l, I)$ describes the relationship between the data and the model. When viewed as a function of the model parameters, the observation model is referred to as the likelihood function, and it plays an important role in statistics where it is mainly used for parameter estimation. However, model comparison cannot only be based on comparing the likelihoods of the candidate models as more complex models can always fit the observed data better than simpler models. In a Bayesian framework, the model parameters and the model order are random variables with the prior pdf $p(\boldsymbol{\theta}_l|l, I)$ and pmf $p(l|I)$, respectively. After observing some data, the state of knowledge is updated by transforming these prior pdfs into the posterior pdfs $p(\boldsymbol{\theta}_l|\boldsymbol{x}, l, I)$ and $p(l|\boldsymbol{x}, I)$ which are connected by Bayes' theorem

$$p(\boldsymbol{\theta}_l|\boldsymbol{x}, l, I) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_l, l, I)p(\boldsymbol{\theta}_l|l, I)}{p(\boldsymbol{x}|l, I)} \tag{B.5}$$

$$p(l|\boldsymbol{x}, I) = \frac{p(\boldsymbol{x}|l, I)p(l|I)}{p(\boldsymbol{x}|I)} \tag{B.6}$$

where

$$p(\boldsymbol{x}|l, I) = \int_{\Theta_l} p(\boldsymbol{x}|\boldsymbol{\theta}_l, l, I)p(\boldsymbol{\theta}_l|l, I)d\boldsymbol{\theta}_l \tag{B.7}$$

is called the marginal likelihood or the evidence. For model comparison, the odds of two competing model orders k and i are often compared. In this connection, the posterior odds are often used, and they are given by

$$\frac{p(k|\boldsymbol{x}, I)}{p(i|\boldsymbol{x}, I)} = \mathrm{BF}[k, i]\frac{p(k|I)}{p(i|I)} \tag{B.8}$$

where the Bayes' factor is

$$\mathrm{BF}[k,i] = \frac{p(\boldsymbol{x}|k,I)}{p(\boldsymbol{x}|i,I)} \ . \tag{B.9}$$

Since the prior and posterior pdfs on the model order are discrete, it is easy to find the posterior odds and the posterior distribution once the Bayes' factors are known. For example, the posterior pmf on the model order is

$$p(l|\boldsymbol{x},I) = \frac{\mathrm{BF}[l;k]p(l|I)}{\sum_{i=1}^{L}\mathrm{BF}[i;k]p(i|I)} \tag{B.10}$$

where the model order k labels some user selected base model which we compare all other models against. Therefore, the main challenge in Bayesian model comparison is to compute the Bayes' factor for competing pairs of models. However, before Bayes' theorem can be used to make inference about the fundamental frequency and the model order in Sec. 4, the prior information I must first be turned into an observation model and prior distributions on the model parameters.

3 A Default Probability Model

As alluded to previously, we are here concerned with the development of an inference scheme which automatically follows from a minimum of prior information I. Thus, a fundamental problem in the inference scheme is to specify a probability model which reflects I. The amount of prior information that is assumed can be stated in the following way.

Assumption 3.1
We are given N data points $\{x(t_n)\}_{n=0}^{N-1}$ from a zero-mean real- or complex-valued signal which has been sampled at the known time instances $\{t_n\}_{n=0}^{N-1}$. The signal is wide-sense stationary (WSS) and consists of a predictable part which is periodic, corrupted by additive noise, and bandlimited to the known angular frequency interval $[\omega_a, \omega_b]$.

For a given application, more prior information may be available which should be included in this assumption. For example in a pitch tracking system, the estimates of the last frame is known, and we might also know something about the correlation structure of the amplitudes of the harmonics and the noise based on, e.g., physical properties. However, we are here concerned with a default and application independent inference scheme so only the information I in Ass. 3.1 is assumed. From I, the observation model and the prior distributions on the model parameters are deduced, and together they constitute the probability model. For the deduction, the principles of maximum entropy and transformation groups is used [38, 39]. Using a few minor approximations, the probability model is later slightly altered so that it becomes more analytically tractable. For

notational convenience, the following vectors and matrix are defined

$$\boldsymbol{x} \triangleq \begin{bmatrix} x(t_0) & \cdots & x(t_{N-1}) \end{bmatrix}^T \tag{B.11}$$

$$\boldsymbol{e} \triangleq \begin{bmatrix} e(t_0) & \cdots & e(t_{N-1}) \end{bmatrix}^T \tag{B.12}$$

$$\boldsymbol{\alpha}_l \triangleq \begin{cases} \begin{bmatrix} \alpha_1 & \cdots & \alpha_l \end{bmatrix}^T, & \boldsymbol{x} \in \mathbb{C}^N \\ \begin{bmatrix} a_1 & \cdots & a_l & b_1 & \cdots & b_l \end{bmatrix}^T, & \boldsymbol{x} \in \mathbb{R}^N \end{cases} \tag{B.13}$$

$$\boldsymbol{z}_i \triangleq \begin{bmatrix} \exp(ji\omega t_0) & \cdots & \exp(ji\omega t_{N-1}) \end{bmatrix}^T \tag{B.14}$$

$$\boldsymbol{Z}_l \triangleq \begin{cases} \begin{bmatrix} \boldsymbol{z}_1 & \cdots & \boldsymbol{z}_l \end{bmatrix}, & \boldsymbol{x} \in \mathbb{C}^N \\ \begin{bmatrix} \mathrm{Re}(\boldsymbol{z}_1) & \cdots & \mathrm{Re}(\boldsymbol{z}_l) & \mathrm{Im}(\boldsymbol{z}_1) & \cdots & \mathrm{Im}(\boldsymbol{z}_l) \end{bmatrix}, & \boldsymbol{x} \in \mathbb{R}^N \end{cases} \tag{B.15}$$

where $(\cdot)^T$ denotes matrix transposition, and $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ take the real and imaginary part, respectively, of a complex number.

3.1 The observation model

In order to deduce the observation model, a model for the non-predictable part or the noise must be selected in (B.1) which in vector notation is given by

$$\boldsymbol{x} = \boldsymbol{Z}_l \boldsymbol{\alpha}_l + \boldsymbol{e} \ . \tag{B.16}$$

That is, which distribution should be selected for the noise vector $\boldsymbol{e}$ given the prior information I? Obviously, the distribution must integrate to one and have zero-mean, and the average power of the noise process must be finite since the signal has been sampled. Thus, the noise variance σ^2 is therefore finite, and the WSS property implies that σ^2 does not change with time. As advocated in [38, 39], the pdf which maximises the entropy under these constraints should be selected, and this pdf is the (complex) normal distribution with density

$$p(\boldsymbol{e}|\sigma^2, I) = \left[r\pi\sigma^2\right]^{-N/r} \exp\left(-\frac{\boldsymbol{e}^H \boldsymbol{e}}{r\sigma^2}\right) \tag{B.17}$$

$$= \begin{cases} \mathcal{CN}(\boldsymbol{e}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}_N) \ , & r = 1 \\ \mathcal{N}(\boldsymbol{e}; \boldsymbol{0}, \sigma^2 \boldsymbol{I}_N) \ , & r = 2 \end{cases} \tag{B.18}$$

where $(\cdot)^H$ denotes conjugate matrix transposition, $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and r is either 1 for $\boldsymbol{x} \in \mathbb{C}^N$ or 2 for $\boldsymbol{x} \in \mathbb{R}^N$. To simplify the notation, the non-standard notation $\mathcal{N}_r(\cdot)$ is used to refer to either the complex normal pdf $\mathcal{CN}(\cdot)$ for $r = 1$ or the real normal pdf $\mathcal{N}(\cdot)$ for $r = 2$. It is important to note that the noise variance σ^2 is a random variable. As opposed to the case where it is simply a fixed and unknown

quantity, the noise distribution marginalised over this random noise variance is able to model noise with heavy tails and is robust towards outliers. In Sec. 3.2, the prior distribution on the noise variance is elicited. Note that (B.18) does not explicitly model any correlation structure in the noise. If prior information about such a structure is available, it should be included in the constraints to enable more accurate estimation results. However, including these constraints lowers the entropy so if nothing is known about a correlation structure, (B.18) is the least informative distribution on the noise since it maximises the entropy and is thus able to capture any correlation structure in the noise [40, 41].

From (B.18), it follows that the observation model is

$$p(\boldsymbol{x}|\boldsymbol{\alpha}_l, \sigma^2, \omega, l, I) = \mathcal{N}_r(\boldsymbol{x}; \boldsymbol{Z}_l\boldsymbol{\alpha}_l, \sigma^2\boldsymbol{I}_N) \ . \tag{B.19}$$

In most of the literature on fundamental frequency estimation, the same observation model is used. However, the derivation presented here facilitates a different interpretation of this model. Namely, when nothing is known about the noise except that it is WSS and has a finite power, the white Gaussian noise assumption is the least informative or most conservative noise distribution.

3.2 The Prior Distributions

When the parametrisation is not given by the problem, the maximum entropy method cannot be used for the elicitation of a default prior distribution [42, Sec. 5.6.2]. For example, the noise variance σ^2 has so far been used in the parametrisation, but the standard deviation σ or the precision parameter $\lambda = \sigma^{-2}$ could have been used instead. Applying the maximum entropy principle to either of these three common representations leads to the unsatisfactory situation that the prior distribution is not invariant under the choice of parametrisations. In order to cope with the different representations, the invariances which the prior distribution must obey are often considered [38, 39]. That is, which transformations of the parameters do not change the prior knowledge? Another useful question to consider is which parameters are logically connected. That is, if the value of one parameter is known, would that change the state of knowledge about the other parameters? Although this is not a necessary question to consider, selecting a representation in which the parameters are not logically connected simplifies the prior elicitation [43, App. A]. In our representation, the parameters are the complex amplitudes $\boldsymbol{\alpha}_l$, the noise variance σ^2, the fundamental frequency ω, and the model order l. The fundamental frequency is clearly logically connected to the model order since it must be below ω_b/l. However, other dependencies between the parameters cannot be extracted from our prior information I, and the prior pdf is therefore factored as

$$p(\boldsymbol{\alpha}_l, \sigma^2, \omega, l|I) = p(\boldsymbol{\alpha}_l|I)p(\sigma^2|I)p(\omega|l, I)p(l|I) \ . \tag{B.20}$$

The Noise Variance

Since the choice of parametrisation is not obvious, the prior distribution on the noise variance is selected such that it does not depend on whether the noise variance, the precision parameter, or the standard deviation is used. For invariance under either of these representations, it is therefore required that

$$p(\sigma|I)d\sigma = p(\sigma^m|I)d\sigma^m \ , \qquad \forall\ m \neq 0 \tag{B.21}$$

which is satisfied for $p(\sigma|I) \propto \sigma^{-1}$. This improper prior pdf is very famous and known as the Jeffreys' prior [44]. It is improper since it does not integrate to one. In practice, however, the noise variance cannot go all the way to zero due to, for example, quantisation noise, and the noise variance is always upper bounded so a normalised prior pdf on the noise variance is

$$p(\sigma^2|I) = \begin{cases} \left[\ln(w/v)\sigma^2\right]^{-1} & v < \sigma^2 < w \\ 0 & \text{otherwise} \end{cases} . \tag{B.22}$$

The bounds on the noise variance have almost no influence on the inference so they are often selected as $v \to 0$ and $w \to \infty$ to simplify the analysis [43, App. A].

The Fundamental Frequency

For the elicitation of the prior distribution on the fundamental frequency, the arguments from Sec. 3.2 can be repeated. Whether the (angular) fundamental frequency ω, the ordinary fundamental frequency $f = \omega/(2\pi)$, or the fundamental period $\tau = f^{-1}$ is used, does not change the prior knowledge. That is, the prior distribution should be invariant under any transformation of the form $k\omega^m$ for any positive constant k. From the prior information I, the signal is bandlimited to the interval $[\omega_a, \omega_b]$. Given the model order l, ω must therefore lie on the interval $\Omega_l = [\omega_a, \omega_b/l]$. Thus, using the same arguments as for the noise variance, the posterior pdf on the fundamental frequency is

$$p(\omega|l, I) = \begin{cases} (F_l\omega)^{-1} & \omega \in \Omega_l \\ 0 & \text{otherwise} \end{cases} \tag{B.23}$$

where $F_l = \ln(\omega_b) - \ln(l\omega_a)$. This prior was also derived in [43, App. A] for a single sinusoid using a more ingenious argument.

The Complex Amplitudes

The sinusoidal model is typically parametrised by the Cartesian coordinates (a_i, b_i) or the polar coordinates (A_i, ϕ_i). Since neither of these representations change the state

of knowledge, the prior pdf on the complex amplitudes is required to be invariant under the transformation between these two representations. That is,

$$p(a_i, b_i|I)da_i db_i = q(A_i, \phi_i|I)A_i dA_i d\phi_i \tag{B.24}$$

for $i = 1, \cdots, l$ where $p(a_i, b_i|I)$ and $q(A_i, \phi_i|I)$ are the pdfs on the Cartesian and polar coordinates, respectively. From the prior information I, the signal is assumed to be zero mean and WSS. In terms of the Cartesian coordinates, this implies that a_i and b_i are uncorrelated and both have zero mean and the same expected power $\sigma_\alpha^2/2$. For the polar coordinates, it implies that the phase is uniformly distributed on any continuous interval of length 2π and uncorrelated with the amplitude. Finally, since the phases $\{\phi_i\}_{i=1}^l$ are independent and uniformly distributed, the l harmonic components are uncorrelated [45, Ch. 4]. We note in passing that many of the same arguments are also used for the derivation of the covariance matrix model for a time series. For the marginal pdfs f and g, (B.24) can therefore be written as

$$f(a_i)f(b_i) = (2\pi)^{-1}g(A_i) \ , \quad i = 1, \cdots, l \ . \tag{B.25}$$

For $a_i = c$ and $b_i = 0$, this reduces to $g(c) = 2\pi f(c)f(0)$ so that (B.25) becomes

$$f(a_i)f(b_i) = f\left(\sqrt{a_i^2 + b_i^2}\right) f(0) \ , \quad i = 1, \cdots, l \ . \tag{B.26}$$

The only possible solution to this functional equation is that f is a normal pdf [39, Ch. 7] so that

$$p(a_i, b_i|\sigma_\alpha^2, I) = \mathcal{N}_2([a_i, b_i]^T; \mathbf{0}, (\sigma_\alpha^2/2)\boldsymbol{I}_2) \ . \tag{B.27}$$

Turning this bivariate real normal pdf into a univariate complex normal pdf on the complex amplitude α_i gives [46, Ch. 15]

$$p(\alpha_i|\sigma_\alpha^2, I) = \mathcal{N}_1(\alpha_i; 0, \sigma_\alpha^2) \ . \tag{B.28}$$

The joint pdf on $\boldsymbol{\alpha}_l$ is therefore

$$p(\boldsymbol{\alpha}_l|\sigma_\alpha^2, I) = \mathcal{N}_r(\boldsymbol{\alpha}_l; \mathbf{0}, (\sigma_\alpha^2/r)\boldsymbol{I}_{rl}) \ . \tag{B.29}$$

The derivation of the normal pdf given above is often called the Herschel-Maxwell derivation [39]. Since σ_α^2 is unknown, this hyperparameter is treated as a random variable. By using the same arguments as for the noise variance, the following hyperprior is obtained

$$p(\sigma_\alpha^2|I) = \begin{cases} \left[\ln(w/v)\sigma_\alpha^2\right]^{-1} & v < \sigma_\alpha^2 < w \\ 0 & \text{otherwise} \end{cases} \ . \tag{B.30}$$

The Model Order

Since the model order is a discrete parameter, the maximum entropy principle can be applied without worrying about the parametrisation. Under the constraint that the prior pmf of the model order must integrate to one, $p(l)$ is the uniform pmf on the set $l \in \{1, 2, \cdots, L\}$. As model orders larger than $\lfloor \omega_b/\omega_a \rfloor$ have zero support, L should not be chosen larger than this value. Note that the model order $l = 0$ is not in the support set since the prior information I states that a predictable part is present in the signal. However, later on, it is discussed how the proposed algorithm can cope with the detection of a predictable part.

3.3 The g-Prior

In the previous sections, the prior information I has been turned into a default probability model. Unfortunately, the prior probability model renders the inference problem analytically intractable. However, if a re-parametrisation and a few minor approximations are made, a prior on the same form as the Zellner's g-prior [47] is obtained, and this prior has some tractable analytical properties [34, 48]. For the re-parametrisation, the power of the i'th harmonic component is written as

$$\frac{\sigma_\alpha^2}{r} = \frac{rg\sigma^2}{N} \quad \Longleftrightarrow \quad g = \frac{N\sigma_\alpha^2}{r^2\sigma^2} = \frac{N\eta}{rl} \tag{B.31}$$

where the signal-to-noise ratio (SNR) is defined as

$$\eta \triangleq \frac{E[|s(t_n)|^2]}{E[|e(t_n)|^2]} = \sum_{i=1}^{l} \frac{\sigma_\alpha^2}{r\sigma^2} = \frac{l\sigma_\alpha^2}{r\sigma^2} \ . \tag{B.32}$$

Thus, g may be interpreted as N/r times the average SNR. Note that although any prior dependency between the complex amplitudes and the noise variance was included in the factorisation in (B.20), the dependency automatically appears through g. As reviewed in [48], the hyperparameter g can be set to a fixed value or treated as a random variable. When g is a random variable, the prior pdf of g can be derived from (B.31), (B.30), and (B.22) to

$$p(g|I) = \begin{cases} \frac{\ln(w/v) - |\ln(r^2 g/N)|}{\ln^2(w/v) g} \ , & g \in \left[\frac{Nv}{r^2 w}, \frac{Nw}{r^2 v}\right] \\ 0 \ , & \text{otherwise} \end{cases} \tag{B.33}$$

which in the limit of $v \to 0$ and $w \to \infty$ reduces to the prior $p(g|I) \propto g^{-1}$ for $g > 0$.

To justify the approximations, which we introduce below, the following assumption is made.

Assumption 3.2
The number of data points N is large enough to justify that $(N/r)(\boldsymbol{Z}_l^H \boldsymbol{Z}_l)^{-1} \approx \boldsymbol{I}_{rl}$.

Ass. 3.2 is often used in connection with sinusoidal frequency estimation to lower the computational complexity of the inference algorithm significantly. It holds for a uniform sampling scheme and for sufficiently random non-uniform sampling schemes, and it stems from that sinusoids are asymptotically orthogonal for any set of distinct frequencies. That is,

$$\lim_{N\to\infty} \frac{r}{N} \boldsymbol{Z}_l^H \boldsymbol{Z}_l = \boldsymbol{I}_{rl} \ . \tag{B.34}$$

For a fixed N, the approximation gets progressively worse as the frequencies become smaller and closer [2]. Under Ass. 3.2 and the re-parametrisation in (B.31), the prior pdf on the complex amplitudes becomes

$$p(\boldsymbol{\alpha}_l|\sigma^2,\omega,g,I) = \mathcal{N}_r(\boldsymbol{\alpha}_l;\boldsymbol{0},g\sigma^2(\boldsymbol{Z}_l^H\boldsymbol{Z}_l)^{-1}) \ . \tag{B.35}$$

Another consequence of Ass. 3.2 is that the likelihood function for the fundamental frequency is very sharply peaked around the ML estimate of ω. Therefore, the prior distribution on ω only has negligibly effect on the posterior distribution [43, App. A], and it is therefore approximated by a uniform pdf on the interval Ω_l. That is,

$$p(\omega|l) = W_l^{-1}\mathbb{I}_{\omega_l}(\omega) \tag{B.36}$$

where $W_l = \omega_b/l - \omega_a$ and $\mathbb{I}_{\omega_l}(\omega)$ is the indicator function on the interval Ω_l.

As noted in Sec. 3.2, the bounds on the noise variance have almost no influence on the inference. They are therefore selected as $v \to 0$ and $w \to \infty$ so that the improper Jeffreys' prior $p(\sigma^2|I) \propto (\sigma^2)^{-1}$ is obtained for the noise variance. For Bayesian comparison of models with parameter spaces of different dimensions, proper prior distributions must be selected on the model parameters to make the Bayes' factor well-defined [34]. However, since the noise variance is a common parameter in all models, an improper prior may be used on it [49]. Since g is also a common parameter in all models, the prior $p(g|I) \propto g^{-1}$ may be used for $g > 0$. For example, this prior has been used in [50]. Although simple, this prior does not allow marginalisation w.r.t. g in the inference step. However, the prior is a limiting case of the beta prime or inverted beta distribution with density

$$p(g|\epsilon,\delta,I) = \frac{(\delta-1)\Gamma(\epsilon+\delta)}{\epsilon\Gamma(\epsilon)\Gamma(\delta)} g^{\epsilon}(1+g)^{-\delta-\epsilon}\mathbb{I}_{\mathbb{R}^+}(g) \tag{B.37}$$

which is proper for $\delta > \epsilon + 1 > 0$. Although this prior pdf enables analytical inference w.r.t. g, the special case for $\epsilon = 0$ is only used in the sequel to keep the results simpler. Moreover, this special case was also suggested in [48], and it involves only the single hyperparameter $\delta > 1$. Since it is proper, it can be used to detect if a predictable part is present. In the limit of $\delta \to 1$, the improper and user-parameter free prior $p(g|I) \propto (1+g)^{-1}$ is obtained, and it has been shown in [51] that the joint prior $p(g,\sigma^2|I) \propto [\sigma^2(1+g)]^{-1}$ is the Jeffreys' prior and the reference prior [52] for a linear regression model. As this improper prior is a special case of the proper prior on g, the algorithm is derived in the next section for the proper prior. This means that the developed algorithm is able to cope with the detection of predictable part.

4 Bayesian Inference

So far, a default probability model has been developed for the estimation problem based on the prior information I. Based on this model, Bayes' theorem is now used to compute the posterior distributions on the quantities of interest which are the fundamental frequency for every candidate model and the model order. In order to cope with the detection of a predictable part in the signal, the proper prior distribution on g is used. The joint posterior pdf on all model parameters and the number of harmonics is[1]

$$\begin{aligned} p(\boldsymbol{\alpha}_l, \sigma^2\omega, g|\boldsymbol{x}, l) &\propto p(\boldsymbol{x}|\boldsymbol{\alpha}_l, \sigma^2, \omega, g, l)p(\boldsymbol{\alpha}_l|\sigma^2, \omega, g, l) \\ &\quad \times p(\sigma^2)p(\omega|l)p(g) \\ &\propto \mathcal{N}_r(\boldsymbol{\alpha}_l; c\hat{\boldsymbol{\alpha}}_l, \sigma^2\boldsymbol{C}_l)\text{Inv-}\mathcal{G}(\sigma^2; N/r, N\hat{\sigma}_l^2/r) \\ &\quad \times \frac{\Gamma(N/r)\mathbb{I}_{\Omega_l}(\omega)\mathbb{I}_{\mathbb{R}^+}(g)}{(\pi N\hat{\sigma}_l^2)^{N/r}W_l(1+g)^{l+\delta}} \end{aligned} \tag{B.38}$$

where Inv-$\mathcal{G}$ is the inverse gamma pdf. Moreover, we have defined

$$\hat{\boldsymbol{\alpha}}_l \triangleq (\boldsymbol{Z}_l^H\boldsymbol{Z}_l)^{-1}\boldsymbol{Z}_l^H\boldsymbol{x} \tag{B.39}$$

$$c \triangleq g(1+g)^{-1} \tag{B.40}$$

$$\boldsymbol{C}_l \triangleq c(\boldsymbol{Z}_l^H\boldsymbol{Z}_l)^{-1} \tag{B.41}$$

$$\hat{\sigma}_l^2 \triangleq \frac{\boldsymbol{x}^H(\boldsymbol{I}_N - c\boldsymbol{P}_l)\boldsymbol{x}}{N} \tag{B.42}$$

$$R_l^2(\omega) \triangleq \frac{\boldsymbol{x}^H\boldsymbol{P}_l\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}}\ . \tag{B.43}$$

When the ML estimate of the fundamental frequency is used and $c = 1$, $\hat{\sigma}_l^2$ is the ML estimate of the noise variance. The matrix $\boldsymbol{P}_l$ is the orthogonal projection matrix onto the space spanned by the columns of $\boldsymbol{Z}_l$, and $R_l^2(\omega)$ resembles the coefficient of determination from linear regression analysis where it is used to measure the prediction performance. Integrating (B.38) over the noise variance and the complex amplitudes gives

$$p(\omega, g|\boldsymbol{x}, l) \propto \frac{m_0(\boldsymbol{x})(\delta-1)f_l(\omega, g, \delta)\mathbb{I}_{\Omega_l}(\omega)\mathbb{I}_{\mathbb{R}^+}(g)}{W_l} \tag{B.44}$$

where

$$m_0(\boldsymbol{x}) \triangleq \Gamma(N/r)(\pi\boldsymbol{x}^H\boldsymbol{x})^{-N/r} \propto p(\boldsymbol{x}|l=0) \tag{B.45}$$

$$f_l(\omega, g, \delta) \triangleq (1+g)^{N/r-l-\delta}[1+g(1-R_l^2(\omega))]^{-N/r} \tag{B.46}$$

[1] To keep the notation uncluttered, the explicit dependence on the prior information I is omitted in the rest of the paper.

are the unnormalised marginal likelihood for the noise-only model and a very important function in the sequel, respectively. In the case where g is a known parameter, the marginal posterior pdf on ω under model order l is proportional to this function

$$p(\omega|\boldsymbol{x}, g, l) = \frac{p(\omega, g, l|\boldsymbol{x})}{p(g)p(l)} \propto f_l(\omega, g, 0)\mathbb{I}_{\Omega_l}(\omega) \ . \tag{B.47}$$

When g is an unknown parameter and $l > 1 - \delta$, it can be integrated out of (B.44) so that the marginal posterior pdf on ω under model order l is obtained as

$$\begin{aligned} p(\omega|\boldsymbol{x}, l) &= \int_0^\infty \frac{p(\omega, g, l|\boldsymbol{x})}{p(l)} dg \propto \int_0^\infty f_l(\omega, g, \delta)\mathbb{I}_{\Omega_l}(\omega) dg \\ &\propto {}_2F_1(N/r, 1; l + \delta; R_l^2(\omega))\mathbb{I}_{\Omega_l}(\omega) \end{aligned} \tag{B.48}$$

where ${}_2F_1$ is the Gaussian hypergeometric function [53, p. 314]. The condition $l > 1 - \delta$ ensures that the integral in (B.48) converges, and it is satisfied for all l when the prior on g is proper, i.e., $\delta > 1$, and for any $l > 0$ even for the improper prior on g with $\delta \to 1$. The marginal pmf on the model order is given by

$$p(l|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|l)p(l)}{p(\boldsymbol{x})} = \frac{\mathrm{BF}[l, 0]p(l)}{\sum_{i=0}^{L} \mathrm{BF}[i, 0]p(i)} \tag{B.49}$$

where $p(\boldsymbol{x}|l, I)$ is the marginal likelihood and

$$\mathrm{BF}[l, 0] = \frac{p(\boldsymbol{x}|l)}{p(\boldsymbol{x}|l = 0)} = \frac{m_l(\boldsymbol{x})}{m_0(\boldsymbol{x})} \tag{B.50}$$

is the Bayes' factor. Here, the noise-only model is used as the base model so the prior distribution on g must be proper. When the noise-only model is not in the set of candidate models, the model with a single harmonic component is used as the base model. In this case, the prior on g can be improper, and the Bayes' factor is given by

$$\mathrm{BF}[l, 1] = \lim_{\delta \to 1} \frac{\mathrm{BF}[l, 0]}{\mathrm{BF}[1, 0]} \ . \tag{B.51}$$

When g is a known parameter, the Bayes' factor has the following integral representation

$$\begin{aligned} \mathrm{BF}[l, 0|g] &= \frac{p(l = 0) \int_{\Omega_l} p(\omega, g, l|\boldsymbol{x}) d\omega}{p(l = 0|\boldsymbol{x})p(g)p(l)} \\ &= \frac{1}{W_l} \int_{\Omega_l} f_l(\omega, g, 0) d\omega \ , \end{aligned} \tag{B.52}$$

and when g is an unknown parameter, the Bayes' factor is

$$\mathrm{BF}[l,0] = \frac{\delta-1}{W_l}\int_{\Omega_l}\int_0^\infty f_l(\omega,g,\delta)dgd\omega \tag{B.53}$$

$$= \frac{\delta-1}{W_l(l+\delta-1)}\int_{\Omega_l} {}_2F_1(N/r,1;l+\delta;R_l^2(\omega))d\omega\ . \tag{B.54}$$

Unfortunately, the modes and the moments of the posterior pdf on the fundamental frequency are not available in closed-form due to the non-linear way that ω parametrises the pdfs in (B.47) and (B.48). Moreover, the posterior model order probabilities are not available in closed-form since the integrals in (B.52) and (B.54) cannot be computed analytically. In Sec. 5, various approximate ways of finding these modes, moments, and posterior probabilities are discussed.

4.1 Selecting a Value for g

In order to facilitate an easier evaluation of the posterior pdfs for the fundamental frequency and the model order, the parameter g is often assumed to be a deterministic parameter rather than a random variable when it is unknown. Thus, instead of marginalising over g, a value for g is selected or estimated. There exist several ways of selecting the value of g, and two popular choices are considered here. Selecting $g^{\mathrm{BIC}} = N$ approximately corresponds to the Bayesian information criterion (BIC) [48]. Alternatively, an empirical Bayesian method can be used in which the unknown hyperparameter g is estimated from the data. The value of g can then be estimated as the maximum likelihood estimate of the joint pdf $p(\boldsymbol{x},\boldsymbol{\alpha}_l,\sigma^2,\omega|g,l)$ integrated w.r.t. the unknown parameters. However, since the marginalisation over the fundamental frequency cannot be done in closed-form, the marginalisation is only carried out over the complex amplitudes and the noise variance, and the fundamental frequency is simply replaced with its MAP estimate $\hat{\omega}$ which is derived in the next section. That is,

$$\begin{aligned} g_l^{\mathrm{EB}} &= \arg\max_{g\in\mathbb{R}^+} p(\boldsymbol{x},\hat{\omega}|g,l) = \arg\max_{g\in\mathbb{R}^+} p(\hat{\omega}|\boldsymbol{x},g,l) \\ &= \arg\max_{g\in\mathbb{R}^+} f_l(\hat{\omega},g,0) = \max\left(\frac{NR_l^2(\hat{\omega})-rl}{(1-R_l^2(\hat{\omega}))rl},0\right)\ . \end{aligned} \tag{B.55}$$

There are several other ways of selecting the value of g, and the interested reader is referred to the excellent review in [48] and the references therein.

5 Approximations

As stated in the previous section, the goal is to find the modes and the variances of the fundamental frequency for every candidate model as well as the posterior model

probabilities. In this section, several approximations for various choices of g are derived. The accuracy of these approximations is evaluated in a small-scale simulation study in Sec. 7.

5.1 Numerical Integration

Since the integrals in (B.52) and (B.54) are one dimensional integrals, they can easily be evaluated using numerical integration techniques. For example, the integrals in (B.52) and (B.54) can be approximately evaluated by computing

$$\mathrm{BF}[l,0|g] \approx \frac{1}{K}\sum_{k=1}^{K} f_l(\omega_k, g, 0) \tag{B.56}$$

$$\mathrm{BF}[l,0] \approx \frac{(\delta-1)}{K(l+\delta-1)}\sum_{k=1}^{K} {}_2F_1(N/r, 1; l+\delta; R_l^2(\omega)) \ , \tag{B.57}$$

respectively, where $\{\omega_k\}_{k=1}^K$ are K equidistant candidate frequencies from the set Ω_l with $W_l/K = \omega_{k+1} - \omega_k$, $\omega_1 = \omega_a$, and $\omega_K = \omega_b/l - W_l/K$. However, the functions $f_l(\omega, g, 0)$ and ${}_2F_1(N/r, 1; l+\delta; R_l^2(\omega))$ are usually very sharply peaked around the mode of the fundamental frequency so the pdfs have to be evaluated over a fine frequency grid to make the approximation accurate. Moreover, the computation of $f_l(\omega_k, g, 0)$ and, in particular, ${}_2F_1(N/r, 1; l+\delta; R_l^2(\omega_k))$ is quite costly since either $\boldsymbol{x}^H\boldsymbol{P}_l\boldsymbol{x}$ or ${}_2F_1$ has to be computed for all K candidate frequencies. Even under Ass. 3.2, the limit in (B.34) cannot be used to justify the approximation

$$\boldsymbol{x}^H\boldsymbol{P}_l\boldsymbol{x} \approx \frac{r}{N}\|\boldsymbol{Z}_l^H\boldsymbol{x}\|^2 \tag{B.58}$$

since the value of $f_l(\omega, g, \delta)$ is very sensitive to even small perturbations in $R_l^2(\omega)$ when it is close to one and the SNR is large. Thus, the numerical integration of (B.52) and (B.54) may entail a too high computational load, and some analytical approximations are therefore also considered since they can reduce this computational load significantly.

5.2 The Distribution on the Fundamental Frequency

Although a closed-form expression has been derived for the pdf of the fundamental frequency for both a known and an unknown g in (B.47) and (B.48), respectively, neither its moments nor its mode can be found in closed-form. The pdf of the fundamental frequency is therefore approximated by a normal pdf whose mean, mode, and variance are easily accessible. This approximation is also useful for the evaluation of the integrals in (B.52) and (B.54). The normal approximation is accurate when the following assumption is true.

Assumption 5.1
The SNR is high enough to justify that the posterior pdfs $p(\omega|\boldsymbol{x}, g, l)$ and $p(\omega|\boldsymbol{x}, l)$ of the fundamental frequency for a known and an unknown g, respectively, consist only of a single important symmetric peak.

Under adverse signal conditions such as a low SNR, Ass. 5.1 is false since the pdfs $p(\omega|\boldsymbol{x}, g, l)$ and $p(\omega|\boldsymbol{x}, l)$ are multi-modal. In this case, the distribution on the fundamental frequency may be approximated by a Gaussian mixture model instead [54, Ch. 12]. However, this is not explored any further in this paper. In Sec. 7, the accuracy of Ass. 5.1 is evaluated. The normal approximation of $p(\omega|\boldsymbol{x}, g, l)$ is

$$p(\omega|\boldsymbol{x}, g, l) \approx \mathcal{N}_2(\omega; \hat{\omega}, s_l(\hat{\omega}|g)) \tag{B.59}$$

where $\hat{\omega}$ is the mode of $p(\omega|\boldsymbol{x}, g, l)$ corresponding to the MAP estimate of the fundamental frequency, and

$$s_l(\hat{\omega}|g) = -\left[\frac{\partial^2 \ln p(\omega|\boldsymbol{x}, g, l)}{\partial \omega^2}\bigg|_{\omega=\hat{\omega}}\right]^{-1} . \tag{B.60}$$

The normal approximation

$$p(\omega|\boldsymbol{x}, l) \approx \mathcal{N}_2(\omega; \hat{\omega}, s_l(\hat{\omega})) \tag{B.61}$$

has the same mean, but the variance is

$$s_l(\hat{\omega}) = -\left[\frac{\partial^2 \ln p(\omega|\boldsymbol{x}, l)}{\partial \omega^2}\bigg|_{\omega=\hat{\omega}}\right]^{-1} . \tag{B.62}$$

As stated above, the MAP estimate of the fundamental frequency under model order l does not depend on whether the value of g is known or not. It is given as the solution to

$$\begin{aligned}\hat{\omega} &= \arg\max_{\omega\in\Omega_l} p(\omega|\boldsymbol{x}, g, l) = \arg\max_{\omega\in\Omega_l} p(\omega|\boldsymbol{x}, l) \\ &= \arg\max_{\omega\in\Omega_l} R_l^2(\omega) = \arg\max_{\omega\in\Omega_l} \boldsymbol{x}^H \boldsymbol{P}_l \boldsymbol{x} ,\end{aligned} \tag{B.63}$$

and it is the same as the ML estimate [45, Ch. 4]. Unfortunately, it is costly from a computational point of view to find the ML estimate since the cost-function in (B.63) has a complicated multi-modal shape and is very sharply peaked around $\hat{\omega}$, especially for a high SNR. Typically, the ML estimate is found by first evaluating the cost-function on a fine grid and then performing a local optimisation around the maximum value of the cost-function on this grid. However, the computational complexity of this procedure may be too high since the projection matrix $\boldsymbol{P}_l$ must be evaluated for every

candidate frequency. The computational cost can be significantly reduced by making the approximation in (B.58). This leads to the following approximate MAP-estimate

$$\hat{\omega} \approx \arg\max_{\omega \in \Omega_l} \boldsymbol{x}^H \boldsymbol{Z}_l \boldsymbol{Z}_l^H \boldsymbol{x} = \arg\max_{\omega \in \Omega_l} \|\boldsymbol{Z}_l^H \boldsymbol{x}\|_2^2 \tag{B.64}$$

which under a uniform sampling frequency can be computed efficiently using a single FFT [2]. To get the MAP estimate in (B.63), the approximate MAP estimate in (B.64) may be used as the starting point of a local optimisation using the exact cost-function in (B.63). The local optimisation can also be substituted for faster and approximate techniques based on, e.g., interpolation [55].

In order to find the variances $s_l(\hat{\omega}|g)$ and $s_l(\hat{\omega})$ of the fundamental frequency, the second order derivatives of $\ln p(\omega|\boldsymbol{x}, g, l)$ and $\ln p(\omega|\boldsymbol{x}, l)$ must be found and evaluated at the mode $\hat{\omega}$. The first order derivatives are given by

$$\frac{\partial \ln p(\omega|\boldsymbol{x}, g, l)}{\partial \omega} = \frac{c}{r\hat{\sigma}_l^2} \frac{\partial C_l(\omega)}{\partial \omega} \tag{B.65}$$

$$\frac{\partial \ln p(\omega|\boldsymbol{x}, l)}{\partial \omega} = \frac{{}_2F_1(N/r+1, 2; l+\delta+1; R_l^2(\omega))}{r\hat{\sigma}_0^2(l+\delta){}_2F_1(N/r, 1; l+\delta; R_l^2(\omega))} \times \frac{\partial C_l(\omega)}{\partial \omega} \tag{B.66}$$

where

$$C_l(\omega) \triangleq \boldsymbol{x}^H \boldsymbol{P}_l \boldsymbol{x} \ . \tag{B.67}$$

Note that for $l = 1$, $C_l(\omega)$ is the periodogram. Evaluated at the mode, the second-order derivatives are

$$\left.\frac{\partial^2 \ln p(\omega|\boldsymbol{x}, g, l)}{\partial \omega^2}\right|_{\omega=\hat{\omega}} = \frac{c}{r\hat{\sigma}_l^2} d_2 \tag{B.68}$$

$$\left.\frac{\partial^2 \ln p(\omega|\boldsymbol{x}, l)}{\partial \omega^2}\right|_{\omega=\hat{\omega}} = \frac{{}_2F_1(N/r+1, 2; l+\delta+1; R_l^2(\hat{\omega}))}{r\hat{\sigma}_0^2(l+\delta){}_2F_1(N/r, 1; l+\delta; R_l^2(\hat{\omega}))} \times d_2 \ . \tag{B.69}$$

where

$$d_2 \triangleq \left.\frac{\partial^2 C_l(\omega)}{\partial \omega^2}\right|_{\omega=\hat{\omega}} \ . \tag{B.70}$$

Both of these second order derivatives consist of the second-order derivative of $C_l(\omega)$. It is given by

$$\begin{aligned} d_2 = 2\mathrm{Re}\left[\hat{\boldsymbol{e}}^H \boldsymbol{D}_2 \hat{\boldsymbol{\alpha}}_l - 2\hat{\boldsymbol{e}}^H \boldsymbol{D}_1 (\boldsymbol{Z}_l^H \boldsymbol{Z}_l)^{-1} \boldsymbol{Z}_l^H \boldsymbol{D}_1 \hat{\boldsymbol{\alpha}}_l\right] \\ + 2\hat{\boldsymbol{e}}^H \boldsymbol{D}_1 (\boldsymbol{Z}_l^H \boldsymbol{Z}_l)^{-1} \boldsymbol{D}_1^H \hat{\boldsymbol{e}} - 2\hat{\boldsymbol{\alpha}}_l^H \boldsymbol{D}_1^H \boldsymbol{P}_l^{\perp} \boldsymbol{D}_1 \hat{\boldsymbol{\alpha}}_l \end{aligned} \tag{B.71}$$

where $\boldsymbol{P}_l^{\perp} = \boldsymbol{I}_N - \boldsymbol{P}_l$ and

$$\hat{\boldsymbol{e}} \triangleq \boldsymbol{x} - \boldsymbol{Z}_l \hat{\boldsymbol{\alpha}}_l \tag{B.72}$$

$$\boldsymbol{D}_1 \triangleq \left.\frac{\partial \boldsymbol{Z}_l}{\partial \omega}\right|_{\omega=\hat{\omega}} = j^r (\boldsymbol{1}_r^T \otimes \boldsymbol{t}\boldsymbol{l}^T) \odot \boldsymbol{Z}_l (\boldsymbol{J}_r \otimes \boldsymbol{I}_l) \tag{B.73}$$

$$\boldsymbol{1}_r \triangleq \begin{cases} 1\,, & r = 1 \\ \begin{bmatrix} 1 & -1 \end{bmatrix}^T\,, & r = 2 \end{cases} \tag{B.74}$$

$$\boldsymbol{D}_2 \triangleq \left.\frac{\partial^2 \boldsymbol{Z}_l}{\partial \omega^2}\right|_{\omega=\hat{\omega}} = -(\boldsymbol{1}_r^T \otimes \boldsymbol{t}\boldsymbol{l}^T) \odot (\boldsymbol{1}_r^T \otimes \boldsymbol{t}\boldsymbol{l}^T) \odot \boldsymbol{Z}_l \tag{B.75}$$

$$\boldsymbol{t} \triangleq \begin{bmatrix} t_0 & t_1 & \cdots & t_{N-1} \end{bmatrix}^T \tag{B.76}$$

$$\boldsymbol{l} \triangleq \begin{bmatrix} 1 & 2 & \cdots & l \end{bmatrix}^T\,. \tag{B.77}$$

The operators $\otimes$ and $\odot$ are the Kronecker and Hadamard products, respectively, and $\boldsymbol{J}_r$ is the $r \times r$ exchange matrix. In order to decrease the computational cost of finding the variance of the fundamental frequency, a simpler, but only approximate, expression for the second-order derivative of $C_l(\omega)$ is also derived. Under Ass. 5.1 and at the mode $\hat{\omega}$, it follows that

$$\|\hat{\boldsymbol{\alpha}}_l\|^2 \gg \|\hat{\boldsymbol{e}}\|^2\,. \tag{B.78}$$

Thus, the second order derivative of $C_l(\omega)$ can be approximated by only the last term in (B.71). That is,

$$d_2 \approx -2\hat{\boldsymbol{\alpha}}_l^H \boldsymbol{D}_1^H \boldsymbol{P}_l^{\perp} \boldsymbol{D}_1 \hat{\boldsymbol{\alpha}}_l\,. \tag{B.79}$$

If the limit in (B.34) is used as an approximation, d_2 reduces to

$$\begin{aligned} d_2 \approx\; & -2\hat{\boldsymbol{\alpha}}_l^H \boldsymbol{D}_1^H \boldsymbol{D}_1 \hat{\boldsymbol{\alpha}}_l + \frac{2r}{N} \hat{\boldsymbol{\alpha}}_l^H \boldsymbol{D}_1^H \boldsymbol{Z}_l \boldsymbol{Z}_l^H \boldsymbol{D}_1 \hat{\boldsymbol{\alpha}}_l \\ \approx\; & -\frac{2}{r} \hat{\boldsymbol{\alpha}}_l^H \operatorname{diag}(\boldsymbol{1}_r \otimes \boldsymbol{l})^2 \hat{\boldsymbol{\alpha}}_l \sum_{n=0}^{N-1} t_n^2 \\ & + \frac{2}{rN} \hat{\boldsymbol{\alpha}}_l^H \operatorname{diag}(\boldsymbol{1}_r \otimes \boldsymbol{l})^2 \hat{\boldsymbol{\alpha}}_l \left[\sum_{n=0}^{N-1} t_n\right]^2 \\ =\; & \frac{2}{r} \sum_{i=1}^{l} \hat{A}_i^2 i^2 \left(\frac{1}{N} \left[\sum_{n=0}^{N-1} t_n\right]^2 - \sum_{n=0}^{N-1} t_n^2 \right) \end{aligned} \tag{B.80}$$

where diag(·) transforms a vector into a diagonal matrix. The second approximation follows from the limits

$$\lim_{N\to\infty} r\boldsymbol{D}_1^H\boldsymbol{Z}_l\left[\sum_{n=0}^{N-1} t_n\right]^{-1} = (-j)^r \mathrm{diag}(\mathbf{1}_r \otimes \boldsymbol{l})(\boldsymbol{J}_r \otimes \boldsymbol{I}_l) \tag{B.81}$$

$$\lim_{N\to\infty} r\boldsymbol{D}_1^H\boldsymbol{D}_1\left[\sum_{n=0}^{N-1} t_n^2\right]^{-1} = \mathrm{diag}(\mathbf{1}_r \otimes \boldsymbol{l})^2 \ . \tag{B.82}$$

Under a uniform sampling frequency with no missing samples, $t_n = nT$ and the second-order derivative of $C_l(\omega)$ at $\hat{\omega}$ can be simplified even further since [46, p. 42]

$$\sum_{n=0}^{N-1} t_n = T\sum_{n=0}^{N-1} n = \frac{TN(N-1)}{2} \tag{B.83}$$

$$\sum_{n=0}^{N-1} t_n^2 = T^2\sum_{n=0}^{N-1} n^2 = \frac{T^2N(N-1)(2N-1)}{6} \ . \tag{B.84}$$

Inserting this into (B.80) leads to the approximation

$$d_2 \approx -\frac{T^2N(N^2-1)}{6r}\sum_{i=1}^{l}\hat{A}_i^2 i^2 \ . \tag{B.85}$$

For a known g, this result has an interesting interpretation since the variance of the fundamental frequency under this approximation is

$$s_l(\hat{\omega}|g) \approx \frac{6r^2\hat{\sigma}_l^2}{cT^2N(N^2-1)\sum_{i=1}^{l}\hat{A}_i^2 i^2} \tag{B.86}$$

which for $c = 1$ is the same as the asymptotic Cramér-Rao lower bound of the fundamental frequency with the true values of the complex amplitudes and the noise variance replaced by their maximum likelihood estimates [11]. For a single real-valued sinusoidal signal, the approximate variance in (B.86) was also derived in [41] using a different approach.

In summary, an exact expression in (B.71) and an approximate expression in (B.80) have been derived for the second-order derivative of $C_l(\omega)$ at $\hat{\omega}$. These expressions are used for computing the variances $s_l(\hat{\omega}|g)$ in (B.60) and $s_l(\hat{\omega})$ in (B.62) of the normal approximation to the pdfs $p(\omega|\boldsymbol{x}, g, l)$ and $p(\omega|\boldsymbol{x}, l)$, respectively. Note that for a known g, the variance of the fundamental frequency is proportional to the estimate $\hat{\sigma}_l^2$ of the noise variance. Thus, the worse the data fit the model, the larger the variance of the estimated fundamental frequency is.

5.3 Model Comparison

By approximating $p(\omega|\boldsymbol{x}, g, l)$ and $p(\omega|\boldsymbol{x}, l)$ by the normal pdfs derived in the previous section, the integrals in (B.52) and (B.54) can be evaluated analytically. An approximation of this form is known as the Laplace approximation. Under the Laplace approximation, the Bayes' factors in (B.52) and (B.54) are

$$\mathrm{BF}[l, 0|g] \approx W_l^{-1} f(\hat{\omega}, g, 0)\sqrt{2\pi s_l(\hat{\omega}|g)} \tag{B.87}$$

$$\mathrm{BF}[l, 0] \approx \frac{(\delta - 1)_2F_1(N/r, 1; l + \delta; R_l^2(\hat{\omega}))\sqrt{2\pi s_l(\hat{\omega})}}{W_l(l + \delta - 1)} \,. \tag{B.88}$$

5.4 The Gaussian Hypergeometric Function

Unfortunately, the Gaussian hypergeometric function is slow to evaluate so from a computational point of view it might not be advantageous to marginalise g analytically in (B.48) and (B.54). Moreover, the use of other priors over g than the hyper-g prior may prohibit analytical marginalisation. Using the Laplace approximation, an approximate way of marginalising (B.48) and (B.54) w.r.t. g is therefore derived [48]. Since the marginal posterior pdf over g is not symmetric and in order to avoid edge effect near $g = 0$, the re-parametrisation $\tau = \ln g$ with the Jacobian $dg/d\tau = \exp(\tau)$ [48] is first made. This re-parametrisation suggest that the posterior distribution over g is approximately a log-normal distribution. With this re-parametrisation, the Laplace approximation of the integral in (B.53) is

$$\begin{aligned}
&\int_{\Omega_l}\int_0^\infty f(\omega, g, \delta)dgd\omega \\
&= \int_{\Omega_l}\int_{-\infty}^\infty \exp(\tau)f(\omega, \exp(\tau), \delta)d\tau d\omega &\text{(B.89)}\\
&= 2\pi\exp(\hat{\tau})f(\hat{\omega}, \exp(\hat{\tau}), \delta)\sqrt{s_l(\hat{\omega}|\exp(\hat{\tau}))\gamma_l(\hat{\tau}|\hat{\omega})} &\text{(B.90)}
\end{aligned}$$

where the mode $\hat{\tau}$ and the variance $\gamma_l(\hat{\tau}|\hat{\omega})$ are given by

$$\hat{\tau} = \ln\left[\frac{\sqrt{\beta_\tau^2 - 4\alpha_\tau} + \beta_\tau}{-2\alpha_\tau}\right] \tag{B.91}$$

$$\begin{aligned}
\gamma_l(\hat{\tau}|\hat{\omega}) = \Bigg[&\frac{N(1 - R_l^2(\hat{\omega}))\exp(\hat{\tau})}{r[1 + \exp(\hat{\tau})(1 - R_l^2(\hat{\omega}))]^2} \\
&- \frac{(N - rl - r\delta)\exp(\hat{\tau})}{r(1 + \exp(\hat{\tau}))^2}\Bigg]^{-1}
\end{aligned} \tag{B.92}$$

with

$$\alpha_\tau \triangleq (1 - R_l^2(\hat{\omega}))(1 - l - \delta) \tag{B.93}$$

$$\beta_\tau \triangleq (N/r - 1)R_l^2(\hat{\omega}) - l - \delta + 2 \; . \tag{B.94}$$

Thus, for $\hat{g} \triangleq \exp(\hat{\tau})$, the Bayes' factor in (B.54) is approximately

$$\mathrm{BF}[l, 0] \approx \frac{2\pi(\delta - 1)\hat{g} f(\hat{\omega}, \hat{g}, \delta)\sqrt{s_l(\hat{\omega}|\hat{g})\gamma_l(\hat{\tau}|\hat{\omega})}}{W_l} \; , \tag{B.95}$$

and the normal approximation of the pdf of the fundamental frequency in (B.48) is approximately

$$p(\omega|\boldsymbol{x}, l) \approx \mathcal{N}_2(\omega; \hat{\omega}, s_l(\hat{\omega}|\hat{g})) \; . \tag{B.96}$$

6 Comparison to an ML Estimator

Before evaluating the proposed inference scheme, it is compared to what we believe is one of the state-of-the-art joint fundamental frequency and model order estimators [2, Sec. 2.6] which is based on the asymptotic MAP rule in [56, 57] and is similar to the rules in, e.g, [25, 26]. Although derived in a ML framework, the method can also be interpreted as an optimal filtering method [21]. Moreover, the same algorithm can be obtained as a special case of one of our approximations based on the BIC model selection rule. As stated earlier, the MAP estimate of the fundamental frequency coincides with the ML estimate of the fundamental frequency. Thus, the proposed point estimator of the fundamental frequency is the same as the suggested point estimate in [2]. However, as we treat the fundamental frequency as a random variable, we have also been able to calculate an approximate variance of the fundamental frequency. For model comparison, [2] does not explicitly work with a Bayes' factor. However, it is easy to rewrite their model order estimator as a Bayes' factor. In our notation, it is given by

$$\mathrm{BF}[l, 0] \approx \frac{(\hat{\sigma}_0^2)^N}{(\hat{\sigma}_l^2|_{c=1})^N \sqrt{N^3} N^l} \; . \tag{B.97}$$

where $\hat{\sigma}_0^2 = \boldsymbol{x}^H \boldsymbol{x}/N$. This Bayes' factor has been derived for complex-valued data using the asymptotic MAP rule proposed in [56, 57]. For a fixed g, a uniform sampling frequency, a complex-valued signal, $T = 1$, and the expression for the variance $s_l(\hat{\omega})$ in (B.86), our expression for the Bayes' factor may be written as

$$\mathrm{BF}[l, 0|g] \approx \frac{(\hat{\sigma}_0^2)^N \sqrt{2\pi}}{(1+g)^l W_l (\hat{\sigma}_l^2)^N} \sqrt{\frac{(1+g)6\hat{\sigma}_l^2}{gN(N^2-1)\sum_{i=1}^l \hat{A}_i^2 i^2}} \; . \tag{B.98}$$

For $g^{\mathrm{BIC}} = N$ and $N \gg 1$, $\mathrm{BF}[l,0|g]$ is

$$\mathrm{BF}[l,0|g] \approx \sqrt{\frac{12\pi}{W_l^2 \sum_{i=1}^{l} \hat{A}_i^2 i^2}} \frac{(\hat{\sigma}_0^2)^N}{(\hat{\sigma}_l^2|_{c=1})^N \sqrt{N^3} N^l} \ . \tag{B.99}$$

Comparing this result with (B.97), it is seen that the model order estimator in [2] implicitly assumes the BIC rule and that

$$\sqrt{\frac{12\pi}{W_l^2 \sum_{i=1}^{l} \hat{A}_i^2 i^2}} \approx 1 \tag{B.100}$$

which is hard to justify in general.

7 Simulations

In this section, the accuracy of the various approximations introduced in Sec. 5 is first evaluated on a synthetic signal. All possible combinations of the approximations are not evaluated, but only the most important ones. These are the various approximations of the posterior pdfs on the fundamental frequency and model order, respectively, for an unknown value of g. Second, the proposed inference scheme is evaluated on a female speech signal.

7.1 Synthetic Signal

To evaluate the accuracy of the various approximations introduced in Sec. 5, Monte Carlo simulations was used for various SNRs. Every Monte Carlo realisation consisted of $N = 100$ data points and was sampled uniformly from a complex-valued, periodic, and synthetic signal. The SNR of the signal was varied in steps of 1 dB from -10 dB to 10 dB, and 500 realisations were generated for every SNR. The fundamental frequency ω was assumed to be smaller than $\omega_b = \pi(lT)^{-1}$ so that the frequency of the highest harmonic component was below the Nyquist frequency. For numerical reasons, ω was also assumed to be larger than $\omega_a = 2\pi(NT)^{-1}$.

An overview over the various approximations are given in Table B.1. In the rows marked with ■, an estimate of g is used whereas g is treated as a random variable in the rows marked with •. The rows marked with ▲ are used for reference and comparison to other algorithms. For the first five rows, either the exact or the approximate expressions can be used for the second order derivative of $C_l(\omega)$ given by (B.71) and (B.80), respectively.

ID	type	$p(\omega\|\boldsymbol{x}, l)$	BF$[l, 1]$	g
$c = 1$	■	(B.59)		∞
BIC	■	(B.59)	(B.87), (B.51)	$N - 1$
EB	■	(B.59)	(B.87), (B.51)	(B.55)
FL	•	(B.59)	(B.95), (B.51)	(B.91)
GHF	•	(B.61)	(B.88), (B.51)	
NI	•		(B.57), (B.51)	
UNI	▲	W_l^{-1}	L^{-1}	
ML	▲		(B.97), (B.51)	
SUB	▲		[2, Sec. 4.6]	

Table B.1: Overview over the various approximations.

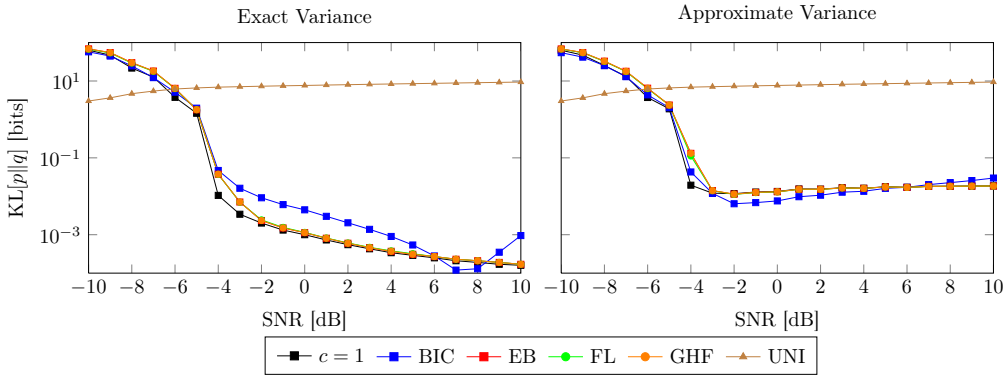

Fig. B.1: The accuracy of the normal approximation of the fundamental frequency under the variance calculations in (B.71) and (B.80), respectively, for various SNRs and choices of g. Note that 'EB', 'FL', and 'GHF' are almost coinciding in the two plots.

The Distribution on the Fundamental Frequency

In order to measure the distance between $p(\omega|\boldsymbol{x}, l)$ and its normal approximation, the relative entropy or Kullback-Leibler (KL) divergence was used. It is given by [58]

$$\mathrm{KL}(p\|q) = \int_{\Omega_l} p(\omega|\boldsymbol{x}, l) \log_2 \left[\frac{p(\omega|\boldsymbol{x}, l)}{q(\omega|\boldsymbol{x}, l)}\right] d\omega \tag{B.101}$$

where $q(\omega|\boldsymbol{x}, l)$ is an approximation of $p(\omega|\boldsymbol{x}, l)$. The KL divergence is finite only if the support of $p(\omega|\boldsymbol{x}, l)$ is contained in Ω_l. Moreover, the KL divergence satisfies that $\mathrm{KL}(p\|q) \geq 0$ with equality if and only if $p(\omega|\boldsymbol{x}, l) = q(\omega|\boldsymbol{x}, l)$. For the true pdf, (B.54) was used, and the KL divergence was evaluated using numerical integration on a fine uniform grid consisting of 10,000 points. Fig. B.1 shows the average KL divergence between $p(\omega|\boldsymbol{x}, l)$ and $q(\omega|\boldsymbol{x}, l)$ for a known model order of $l = 4$. The normal approximation is clearly inaccurate for low SNRs. When the SNR is increased, the approximation becomes better. Below an SNR of approximately -4 dB, the KL divergence is insensitive to the choice of the variance for the normal approximation. However, above -4 dB, the choice matters. For the approximate variance, the KL divergence seems to exhibit a thresholding effect caused by the use of the approximations in (B.58), (B.81), and (B.82). This threshold will be lowered if N is increased.

Model Comparison

In order to evaluate the accuracy of the posterior pmf on the model order, we used the same procedure as in the previous section. Moreover, the model selection properties of the proposed inference scheme was also evaluated and compared to the ML-based algorithm in [2, Sec. 2.6] and the subspace-based algorithm in [2, Sec. 4.6]. The discrete version of the KL divergence is given by [58]

$$\mathrm{KL}(p\|q) = \sum_{l=1}^{L} p(l|\boldsymbol{x}) \ln \left[\frac{p(l|\boldsymbol{x})}{q(l|\boldsymbol{x})}\right] , \tag{B.102}$$

and it is used to assess the accuracy of the posterior pmf $q(l|\boldsymbol{x})$ on the model order for the various approximations in Table B.1. For the true pmf $p(l|\boldsymbol{x})$, the 'NI' approximation based on the numerical integration on a very fine frequency grid was used. For the prior pmf $p(l)$ over the model order, a uniform prior was used so that the posterior pmf on the model order is proportional to the Bayes' factor. The same Monte Carlo simulation setup as above was used but with an unknown model order. Specifically, for each Monte Carlo run, the model order was generated from its prior with the minimum and maximum model order being 1 and $L = 10$, respectively. Since the all-noise model was not in the set of candidate models, the improper prior was used on g, and it is obtained by letting $\delta = 1$. The top row of Fig. B.2 shows the results of measuring the

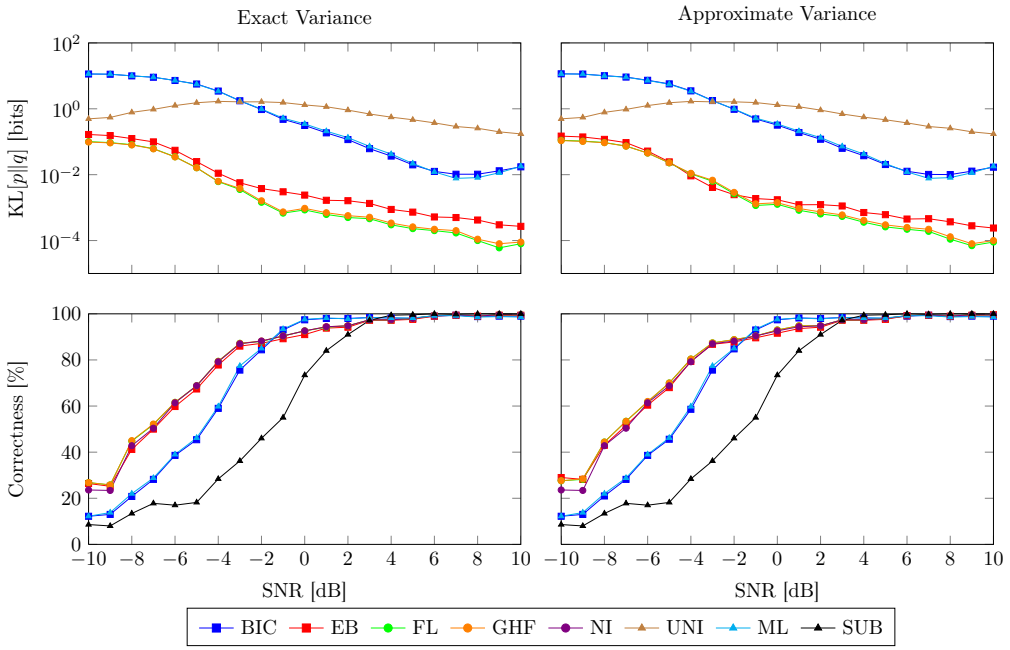

Fig. B.2: The accuracy of the various approximation of the posterior pdf on the model order under the variance calculations in (B.71) and (B.80), respectively, for various SNRs and choices of g. Note that the curve labelled 'UNI' and the curves labelled 'ML' and 'SUB' are only in the plots in the top and bottom row, respectively.

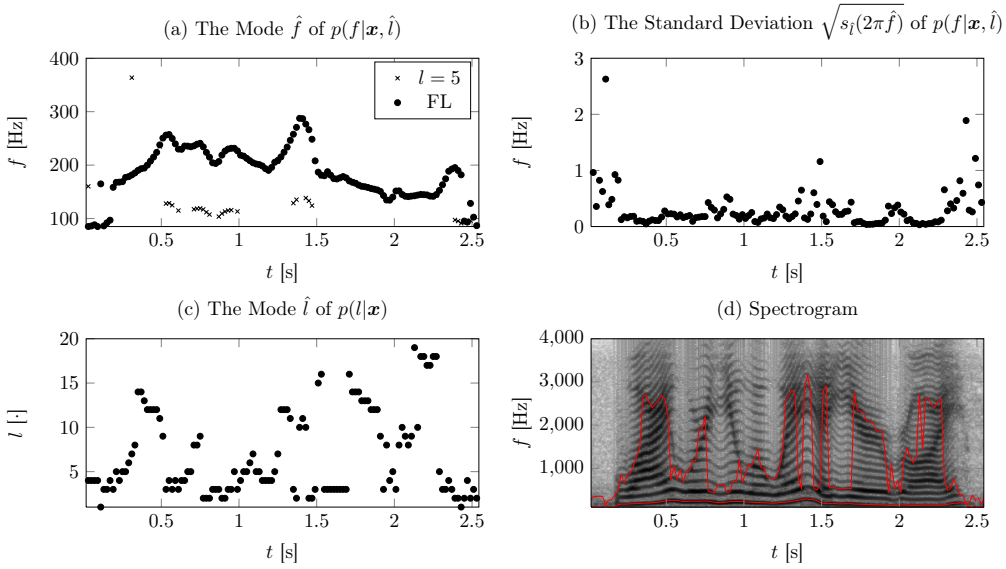

Fig. B.3: The estimation of the fundamental frequency and the model order for a speech signal. Plot (a) and (b) show the estimated fundamental frequency $\hat{f}$ and its standard deviation, respectively, for the estimated model order $\hat{l}$ which is shown in plot (c). Plot (d) shows the spectrogram of the speech signal. The spectrogram has been overlaid with the estimated frequencies for the fundamental and largest harmonic components, respectively.

average KL divergence between $p(l|\boldsymbol{x})$ and $q(l|\boldsymbol{x})$. For all SNRs, the full Laplace 'FL' and the 'GHF' approximations performs slightly better than the approximation based on the emperical bayes 'EB' estimate of g. All of these three approximations perform much better than the 'ML' and the 'BIC' approximations. As shown in Sec. 6, the 'ML' approximations is a special case of the 'BIC' approximation which explains why the 'ML' and the 'BIC' approximations seem to have the same accuracy. In each Monte Carlo run, the most probable model was selected and compared to the true model, and the bottom row of Fig. B.2 shows the proportion of correctly selected model orders for the various SNRs. For SNRs below -2 dB, the 'FL', 'GHF', and 'NI' approximations were better than the other approximations. However, from -2 dB to approximately 3 dB, the 'ML' and 'BIC' approximations were slightly better at finding the true model order. For an SNR above 3 dB, all of the models performed equally well.

Thus, for model selection purposes, there is no best method for all SNRs. However, for problems such as model averaging and estimation in which all models are used, the approximations based on a random g seem to outperform the other approximations for all SNRs.

7.2 Speech Signal

In the last simulation, the applicability of the proposed algorithm was demonstrated to the problem of estimating the fundamental frequency and model order of a speech signal. The speech signal originates from a female voice uttering *"Why were you away a year, Roy?"* which has been sampled at a uniform sampling frequency of 8 kHz. Since the signal is real, the down-sampled analytic signal was first computed as described in the introduction. Subsequently, the signal was partitioned into consecutive frames of 20 ms corresponding to $N = 80$ samples. The minimum and maximum candidate model order were set to 1 and $L = 20$, respectively, and the bandwidth of the signal was set to the interval $[85\text{ Hz}, 4000\text{ Hz}]$ where the lower limit is the typical lower limit of human voiced speech [59, Ch. 6]. For the estimation of the fundamental frequency, the approximate MAP estimate was first estimated using (B.64). Second, a refined estimate was found using a Dichotomous search with the exact cost-function in (B.63). The posterior pmf for the model order was estimated using the 'FL' approximation (see Table B.1) with the approximate variance in (B.86). We have found that the above algorithm provides a good balance between computational load and estimation accuracy[2]. The results of running the algorithm is shown in Fig. B.3. Plot (a) and (b) show the MAP estimate and the standard deviation, respectively, of the fundamental frequency for the estimated model order which is shown in plot (c). In plot (a), the estimated fundamental frequency is also shown for a fixed model order of $l = 5$. We clearly see that the estimator based on a fixed model order suffers from pitch halving, and this illustrates why model order selection is important even if only the estimate of the fundamental frequency is interesting. In plot (d), the frequencies of the fundamental and largest harmonic components are shown on top of the spectrogram of the speech signal. We clearly see that the algorithm provided accurate estimates of the fundamental frequency and the model order even though the signal is not perfectly periodic.

8 Conclusion

In the first part of this paper, we have argued for and derived a default probability model for both a real- and complex-valued periodic signal in additive noise. Using Jaynes' principles of maximum entropy and transformation groups, the scanty prior information in Ass. 3.1 was turned into an observation model and prior distributions on the model parameters. Subsequently, the prior distributions were turned into a more convenient prior of the same form as the g-prior using a few minor approximations on the signal-to-noise-ratio (SNR) and the number of observations. The g-prior is parametrised by the parameter g which is very important for performing model comparison. Several ways of estimating a value for it was given, and it was also treated as a random variable.

[2] A Matlab implementation of the algorithm is available at http://kom.aau.dk/~jkn/publications/publications.php

In the second part of this paper, a closed-form posterior distribution was derived for the fundamental frequency, and an integral representation of the posterior distributions on the model order was derived for both a known and an unknown value of g. Several approximations to these posterior distributions was also suggested, and it was shown that the state-of-the-art ML estimator is a special case of the approximation based on the Bayesian information criterion.

In the last part of this paper, the various approximations were compared in the simulation section on a synthetic signal. The simulations indicated that the value of g is not important for the posterior distribution on the fundamental frequency. For model comparison, the value of g was very important, and the most accurate approximations was obtained when g was treated as a random variable. The BIC approximation is much worse than the other approximations. For model selection, however, the BIC approximation performed slightly better than the other approximations for an SNR larger than approximately -2 dB. It was also demonstrated that one of the approximations was able to accurately estimate the fundamental frequency and model order of a voiced speech segment which was not perfectly periodic.

References

[1] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments*, 2nd ed. Springer, Jun. 1998.

[2] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, B. H. Juang, Ed. Morgan & Claypool, 2009.

[3] H. Dudley, "The carrier nature of speech," *Bell System Technical Journal*, vol. 19, no. 4, pp. 495–515, Oct. 1940.

[4] R. J. Sluijter, "The development of speech coding and the first standard coder for public mobile telephony," Ph.D. dissertation, Techniche Universiteit Eindhoven, 2005.

[5] G. L. Ogden, L. M. Zurk, M. E. Jones, and M. E. Peterson, "Extraction of small boat harmonic signatures from passive sonar," *J. Acoust. Soc. Am.*, vol. 129, no. 6, pp. 3768–3776, Jun. 2011.

[6] V. K. Murthy, L. J. Haywood, J. Richardson, R. Kalaba, S. Salzberg, G. Harvey, and D. Vereeke, "Analysis of power spectral densities of electrocardiograms," *Mathematical Biosciences*, vol. 12, no. 1–2, pp. 41–51, Oct. 1971.

[7] J. Neuberg, R. Luckett, B. Baptie, and K. Olsen, "Models of tremor and low-frequency earthquake swarms on Montserrat," *J. Volcanology and Geothermal Research*, vol. 101, no. 1–2, pp. 83–104, Aug. 2000.

[8] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2498–2517, Apr. 2006.

[9] S. L. Marple, Jr., "Computing the discrete-time "analytic" signal via FFT," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2600–2603, Sep. 1999.

[10] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," 2012, unpublished manuscript.

[11] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1635–1644, Jul. 2007.

[12] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[13] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 1, pp. 24–33, Feb. 1977.

[14] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

[15] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*. Springer, Apr. 1983.

[16] B. Gold and L. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Am.*, vol. 46, no. 2B, pp. 442–448, Aug. 1969.

[17] B. G. Quinn and P. J. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78, no. 1, pp. 65–74, Mar. 1991.

[18] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Processing*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[19] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation using harmonic music," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, Nov. 2006, pp. 521–524.

[20] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1124–1138, Oct. 1986.

[21] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Advances in Signal Process.*, vol. 13, Jun. 2011.

[22] S. J. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2002, pp. 1769–1772.

[23] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[24] P. Stoica and Y. Selén, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[25] B. G. Quinn, "Estimating the number of terms in a sinusoidal regression," *J. of Time Series Analysis*, vol. 10, no. 1, pp. 71–75, Jan. 1989.

[26] L. Kavalieris and E. J. Hannan, "Determining the number of terms in a trigonometric regression," *J. of Time Series Analysis*, vol. 15, no. 6, pp. 613–625, Nov. 1994.

[27] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 387–392, Apr. 1985.

[28] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 450–458, Feb. 2006.

[29] J.-M. Papy, L. De Lathauwer, and S. Van Huffel, "A shift invariance-based order-selection technique for exponential data modelling," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 473–476, Jul. 2007.

[30] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. on Advances in Signal Process.*, 2009.

[31] P. Stoica, Y. Selén, and J. Li, "Multi-model approach to model selection," *Digital Signal Process.*, vol. 14, no. 5, pp. 399–412, Sep. 2004.

[32] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Jun. 2002.

[33] J. O. Berger and L. R. Pericchi, "The intrinsic Bayes factor for model selection and prediction," *J. Amer. Statistical Assoc.*, vol. 91, no. 433, pp. 109–122, Mar. 1996.

[34] ——, "Objective Bayesian methods for model selection: Introduction and comparison," *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, vol. 38, pp. 135–207, 2001.

[35] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer-Verlag New York, Inc., Jul. 2004.

[36] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier, 1995, vol. 495, ch. 14, pp. 495–518.

[37] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc. Interspeech*, 2005, pp. 537–540.

[38] E. T. Jaynes, "Prior probabilities," *IEEE Trans. Syst. Sci. Cybern.*, vol. 4, no. 3, pp. 227–241, 1968.

[39] ——, *Probability Theory : The Logic of Science*, G. L. Bretthorst, Ed. Cambridge University Press, Apr. 2003.

[40] G. L. Bretthorst, "The near-irrelevance of sampling frequency distributions," in *Max. Entropy and Bayesian Methods*, 1999, pp. 21–46.

[41] E. T. Jaynes, "Bayesian spectrum and chirp analysis," in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G. J. Erickson, Eds. D. Reidel, Dordrecht-Holland, 1987, pp. 1–37.

[42] J. M. Bernardo and A. Smith, *Bayesian Theory*, 1st ed. John Wiley and Sons Ltd, 1994.

[43] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag, Berlin Heidelberg, 1988.

[44] H. Jeffreys, *Theory of Probability*. Oxford University Press, 1939.

[45] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Prentice Hall, May 2005.

[46] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall PTR, Mar. 1993.

[47] A. Zellner, "On assessing prior distributions and Bayesian regression analysis with g-prior distributions," in *Bayesian Inference and Decision Techniques*. Elsevier, 1986.

[48] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, "Mixtures of g priors for Bayesian variable selection," *J. Amer. Statistical Assoc.*, vol. 103, pp. 410–423, Mar. 2008.

[49] R. Strachan and H. K. v. Dijk, "Improper priors with well defined Bayes' factors," Department of Economics, University of Leicester, Discussion Papers in Economics 05/4, 2005.

[50] J. M. Marin and C. P. Robert, *Bayesian core: a practical approach to computational Bayesian statistics*, 1st ed. Springer, Feb. 2007.

[51] R. Guo and P. L. Speckman, "Bayes factor consistency in linear models," in *The 2009 International Workshop on Objective Bayes Methodology*, Jun. 2009.

[52] J. O. Berger, J. M. Bernardo, and D. Sun, "The formal definition of reference priors," *Ann. Stat.*, vol. 37, no. 2, pp. 905–938, Apr. 2009.

[53] I. S. Gradshteĭn, I. M. Ryzhik, and A. Jeffrey, *Table of Integrals, Series, and Products*. Academic Press, 2000.

[54] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Jul. 2003.

[55] M. D. Macleod, "Fast nearly ML estimation of the parameters of real or complex single tones or resolved multiple tones," *IEEE Trans. Signal Process.*, vol. 46, no. 1, pp. 141–148, Jan. 1998.

[56] P. M. Djuric, "A model selection rule for sinusoids in white Gaussian noise," *IEEE Trans. Signal Process.*, vol. 44, no. 7, pp. 1744–1751, Jul. 1996.

[57] ——, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[58] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition*. Wiley-Interscience, Jul. 2006.

[59] R. J. Baken and R. F. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. Singular, 1999.

Paper C

Bayesian Interpolation and Parameter Estimation in a Dynamic Sinusoidal Model

Jesper Kjær Nielsen, Mads Græsbøll Christensen, Ali Taylan Cemgil, Simon John Godsill, and Søren Holdt Jensen

The paper has been published in the
IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19(7), pp. 1986–1998, Sep. 2011.

The layout has been revised.

Abstract

In this paper, we propose a method for restoring the missing or corrupted observations of non-stationary sinusoidal signals which are often encountered in music and speech applications. To model non-stationary signals, we use a time-varying sinusoidal model which is obtained by extending the static sinusoidal model into a dynamic sinusoidal model. In this model, the in-phase and quadrature components of the sinusoids are modelled as first-order Gauss-Markov processes. The inference scheme for the model parameters and missing observations is formulated in a Bayesian framework and is based on a Markov chain Monte Carlo method known as Gibbs sampler. We focus on the parameter estimation in the dynamic sinusoidal model since this constitutes the core of model-based interpolation. In the simulations, we first investigate the applicability of the model and then demonstrate the inference scheme by applying it to the restoration of lost audio packets on a packet-based network. The results show that the proposed method is a reasonable inference scheme for estimating unknown signal parameters and interpolating gaps consisting of missing/corrupted signal segments.

1 Introduction

The interpolation of missing, corrupted and future signal samples is an important task in several applications. For example, speech and audio signals are often transmitted over packet-based networks in which packets may be lost, delayed or corrupted. If the contents of neighbouring packets are correlated, the erroneous packets can be approximately reconstructed by using suitable interpolation techniques. The simplest interpolation techniques employ signal repetition [1] and signal stretching [2]. More advanced interpolation techniques are based on filter bank methods such as GAPES and MAPES [3, 4] or based on signal models such as autoregressive models [5, 6], hidden Markov models [7], and sinusoidal models [8–10]. An integral part of the techniques based on signal modelling is the estimation of the signal parameters. Given estimates of these parameters, signal samples are interpolated by simulating data from the model.

Within the applied speech and audio processing field, the sinusoidal signal model is one of the more popular parametric signal models because voiced speech and signals originating from several musical instruments can be accurately modelled as a sum of sinusoids [11]. In this paper, we initially consider the dampened sinusoidal signal model in its real form given by

$$x_n = \sum_{l=1}^{L} \rho_l^n \left[i_l \cos(\omega_l n) + q_l \sin(\omega_l n) \right] + w_n \tag{C.1}$$

where the sampling indices $n = 1, \ldots, N$ label the uniform sampled data. In the model, i_l, q_l, $\omega_l \in [0, \pi]$ and $\rho_l > 0$ denote the undampened in-phase component, the undamp-

ened quadrature component, the (angular) frequency, and the damping coefficient of the l'th sinusoid, respectively. The observed sample x_n at time index n is the sum of L such dampened sinusoids and a white Gaussian noise term w_n with variance σ_w^2. The model in (C.1) is also sometimes written in its polar form given by

$$x_n = \sum_{l=1}^{L} \rho_l^n \alpha_l \cos(\omega_l n - \varphi_l) + w_n \tag{C.2}$$

where $\alpha_l = \sqrt{i_l^2 + q_l^2}$ and $\varphi_l = \arctan(q_l/i_l)$ are the undampened amplitude and phase of the l'th sinusoid, respectively. In this paper, we refer to the models in (C.1) and (C.2) as static sinusoidal models. This naming convention is adopted in order to distinguish it from the dynamic sinusoidal model, which we introduce later.

The static sinusoidal model and its variations have been subject to extensive research for many years. This is primarily due to the large-scale applicability of the model, and because frequency parameters and damping coefficients enter the model in a non-linear fashion. The latter complicates the estimation problem significantly and several methods for solving this problem have therefore been devised. Most of these estimators are aimed at estimating the frequency parameters. Well-known estimators comprise the Min-Norm method [12], non-linear least squares estimators [13, 14], and the high-order Yule-Walker method [15]. Other well-known estimators are the subspace-based methods such as MUSIC [16], root-MUSIC [17], ESPRIT [18], and weighted subspace fitting [19]. A thorough review of most of these estimators is given in [20]. The theoretical foundation of these estimators is based on classical statistics which is also known as frequentist or orthodox statistics. The other major approach to statistics is Bayesian statistics which offers some conceptual advantages to classical statistics (see, e.g., [21] and [22]). For instance, the Bayesian approach copes with nuisance parameters and signal interpolation in a highly standardised way. However, the history of Bayesian frequency estimators is much shorter because the Bayesian methods often struggle with practical problems such as the evaluation of high-dimensional and intractable integrals. In recent years, various developments in Markov chain Monto Carlo (MCMC) methods (see, e.g., [23]) have largely overcome these problems. Nevertheless, the methods still suffer from a high computational complexity.

Bayesian frequency estimation was first considered by Jaynes and Bretthorst in [24] and [25], respectively. In the pioneering work of the latter, the existence of analytical solutions to the Bayesian frequency estimation problem was demonstrated only in the case of a few sinusoids. Moreover, the general inference problem with multiple sinusoids was solved using suitable analytical approximations, under the assumptions that the sinusoids were well-separated and enough data were available. This was not assumed in [26] and [27] in which the general frequency estimation problem was solved by use of an approximate MCMC technique which led to improved performance for closely spaced sinusoids. The performance was improved even further by Andrieu and Doucet in [28],

where the case of unknown model orders was also considered and solved using reversible jump MCMC [29]. Recently, this work has been extended to the case of complex and dampened sinusoidal signals in [30]. In [31], Bayesian inference in the sinusoidal model was applied to the analysis of western tonal music.

In the static model in (C.2), the undampened amplitude α_l and the phase φ_l are assumed to be constant over a segment of N samples. Although this model is widely applicable, the model assumption violates the behaviour of many real world tonal signals. To better model these signals, the model in (C.2) has been modified in various ways. Typical modifications comprise amplitude and/or phase modulation [32, 33], the representation of the amplitudes and/or phases as a linear combination of atoms from a suitable basis [34], and autoregressive (AR) frequency parameters [10]. In this paper, we use a dynamic sinusoidal model formulation in which the in-phase and quadrature components in (C.1) evolve as a first-order Gauss-Markov process. Within the field of econometrics, this class of dynamic models is referred to as stochastic cyclical models [35, 36]. Two slightly different stochastic cyclical models were given a fully Bayesian treatment using MCMC inference techniques in [37] and [38]. Independently, Cemgil et al. introduced a dynamic sinusoidal model for the application of polyphonic music transcription in [39–41]. In this model, the frequency parameters were discrete random variables, and significant attention was given to the problem of estimating note onset and offset. In the more recent papers [42, 43], Bayesian inference schemes for dynamic sinusoidal models were also considered. Like the proposed inference scheme by Cemgil et al., they base their inference schemes on analytical approximations.

In this paper, we first analyse the dynamic model and discuss its interpretation. In this connection, we show that the in-phase and quadrature components of the dynamic sinusoidal model evolve as first-order Gauss-Markov processes. We also extend the cited work in the previous paragraph by developing an inference scheme for the dynamic sinusoidal model on basis of MCMC inference techniques. Moreover, we consider the more general case in which the frequency parameters are continuous random variables and some of the observations are missing. To achieve this, we develop a Gibbs sampler whose output can be used to form the histograms of the unknown parameters of interest. These histograms have the desirable property that they converge towards the probability distribution of these unknown parameters when the number of generated samples is increased, enabling us to extract statistical features for the model parameters and to perform the interpolation of the missing observations.

The paper is organised as follows. In Sec. 2, we present and analyse the dynamic sinusoidal model. We set up the Bayesian framework for the model in Sec. 3, and the proposed inference scheme based on a Gibbs sampler is derived in Sec. 4. Four simulations are performed in Sec. 5 illustrating the applicability of the model as well as the performance of the sampler and interpolator, and Sec. 6 concludes this paper. The Appendix contains a list of the relevant probability distributions.

2 Dynamic Signal Model

In the static sinusoidal model in (C.1), the undampened in-phase and quadrature components are constant throughout the segment of N samples. In the dynamic sinusoidal model, however, this restriction is no longer imposed. Similar to, e.g., [38, 39], we consider a dynamic sinusoidal model given by

$$\begin{aligned} y_n &= \boldsymbol{b}^T \boldsymbol{s}_n + w_n \qquad \text{(observation equation)} \\ \boldsymbol{s}_{n+1} &= \boldsymbol{A}\boldsymbol{s}_n + \boldsymbol{v}_n \qquad \text{(state equation)} \end{aligned} \tag{C.3}$$

where $\boldsymbol{s}_n$ is a state vector and $\boldsymbol{v}_n$ is a zero-mean Gaussian state noise[1] vector with covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{v}} = \operatorname{diag}(\sigma^2_{\boldsymbol{v},1}\boldsymbol{I}_2, \ldots, \sigma^2_{\boldsymbol{v},l}\boldsymbol{I}_2, \ldots, \sigma^2_{\boldsymbol{v},L}\boldsymbol{I}_2) \ . \tag{C.4}$$

The state noise vectors are mutually independent and independent of the observation noise. Furthermore, we have that

$$\boldsymbol{b} = \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \end{bmatrix}^T \tag{C.5}$$

$$\boldsymbol{A} = \operatorname{diag}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_l, \ldots, \boldsymbol{A}_L) \tag{C.6}$$

$$\boldsymbol{A}_l = \rho_l \begin{bmatrix} \cos\omega_l & \sin\omega_l \\ -\sin\omega_l & \cos\omega_l \end{bmatrix} \ . \tag{C.7}$$

Notice that the state equation of (C.3) decouples into L independent state equations of the form

$$\boldsymbol{s}_{n+1,l} = \boldsymbol{A}_l \boldsymbol{s}_{n,l} + \boldsymbol{v}_{n,l} \tag{C.8}$$

due to the block-diagonal structure of $\boldsymbol{A}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$,

The dynamic model reduces to the static model if there is no state noise. For non-zero state noise, however, the dynamic model models the in-phase and quadrature components as first-order Gauss-Markov processes. In order to see this, we recursively insert the state equation into the observation equation and obtain

$$y_n = \boldsymbol{b}^T \boldsymbol{s}_n + w_n \tag{C.9}$$

$$= \boldsymbol{b}^T \left(\boldsymbol{A}\boldsymbol{s}_{n-1} + \boldsymbol{v}_{n-1}\right) + w_n \tag{C.10}$$

$$= \boldsymbol{b}^T \boldsymbol{A}^n \left(\boldsymbol{A}^{-1}\boldsymbol{s}_1 + \boldsymbol{A}^{-n} \sum_{k=1}^{n-1} \boldsymbol{A}^{k-1} \boldsymbol{v}_{n-k} \right) + w_n \tag{C.11}$$

$$= \sum_{l=1}^{L} \left[i_{n,l} \cos(\omega_l n) + q_{n,l} \sin(\omega_l n) \right] + w_n \tag{C.12}$$

[1] In this paper, noise is not an unwanted component but a random process of interest. We use the term noise for w_n and $\boldsymbol{v}_n$ since this is common practice when working with the state space model in (C.3).

where we have defined

$$\begin{bmatrix} i_{n,l} \\ q_{n,l} \end{bmatrix} \triangleq \rho_l^n \left(\boldsymbol{A}_l^{-1} \boldsymbol{s}_{1,l} + \boldsymbol{A}_l^{-n} \sum_{k=1}^{n-1} \boldsymbol{A}_l^{k-1} \boldsymbol{v}_{n-k,l} \right) . \tag{C.13}$$

Eq. (C.12) is of the same form as (C.1) with one important difference: The in-phase and quadrature components are now time-varying which means that the amplitude and the phases of the polar form of (C.12) are also time-varying. We analyse the statistical behaviour of the time-varying in-phase and quadrature components by introducing the stochastic process defined by $\boldsymbol{z}_{n,l} \triangleq [i_{n,l} \; q_{n,l}]^T$. First, we write $\boldsymbol{z}_{n,l}$ for $n = 1, \ldots, N$ in a recursive way given as

$$\boldsymbol{z}_{n+1,l} = \rho_l \boldsymbol{z}_{n,l} + \left(\rho_l^{-1} \boldsymbol{A}_l \right)^{-(n+1)} \boldsymbol{v}_{n,l} \tag{C.14}$$

with $\boldsymbol{z}_{1,l} = \rho_l \boldsymbol{A}_l^{-1} \boldsymbol{s}_{1,l}$. If we select a Gaussian distribution for the initial state, i.e., $\boldsymbol{s}_{1,l} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{s}_{1,l}}, \sigma^2_{\boldsymbol{s}_{1,l}} \boldsymbol{I}_2)$, then $\boldsymbol{z}_{n,l}$ is a first-order Gauss-Markov process. Second, we notice that, the transformation $\tilde{\boldsymbol{v}}_{n,l} = \left(\rho_l^{-1} \boldsymbol{A}_l \right)^{-(n+1)} \boldsymbol{v}_{n,l}$ is an orthogonal transformation, and we therefore have that

$$p(\tilde{\boldsymbol{v}}_{n,l}) = p(\boldsymbol{v}_{n,l}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{v},l} \boldsymbol{I}_2) . \tag{C.15}$$

Thus, the statistical behaviour of

$$\boldsymbol{z}_{n+1,l} = \rho_l \boldsymbol{z}_{n,l} + \tilde{\boldsymbol{v}}_{n,l} \tag{C.16}$$

is the same as that of (C.14). Therefore, $\boldsymbol{z}_{n,l}$ is a very simple first-order Gauss-Markov process evolving independently of the frequency parameter. Further, if we select the mean and variance of the initial state to be $\boldsymbol{\mu}_{\boldsymbol{s}_{1,l}} = \boldsymbol{0}$ and $\sigma^2_{\boldsymbol{s}_{1,l}} = \sigma^2_{\boldsymbol{v},l}/(1 - \rho_l^2)$, respectively, $\boldsymbol{z}_{n,l}$ is a stationary first-order Gauss-Markov process, i.e., a first order autoregressive process (AR). Also, our model for the observations in (C.3) reduces to a simple AR(1) process if $\omega_l = 0$. In summary, the statistical behaviour of the dynamic model in (C.3) is equivalent to that of the model given by[2]

$$\begin{aligned} \tilde{y}_n &= \sum_{l=1}^{L} \begin{bmatrix} \cos(\omega n) & \sin(\omega n) \end{bmatrix} \begin{bmatrix} \tilde{i}_{n,l} \\ \tilde{q}_{n,l} \end{bmatrix} + w_n \\ \begin{bmatrix} \tilde{i}_{n+1,l} \\ \tilde{q}_{n+1,l} \end{bmatrix} &= \rho_l \begin{bmatrix} \tilde{i}_{n,l} \\ \tilde{q}_{n,l} \end{bmatrix} + \boldsymbol{v}_{n,l} . \end{aligned} \tag{C.17}$$

in which the in-phase and quadrature components are explicitly evolving as a first order Gauss-Markov process. In the model in (C.3), however, the frequencies have been

[2]Here, we have introduced $\tilde{\cdot}$ meaning that, e.g., $\tilde{y}_n \neq y_n$ for the same noise realisations although they share the same statistical behaviour.

separated from the time indices. This makes the inference problem for the frequencies more tractable.

We have shown that the in-phase and quadrature components are modelled as first order Gauss-Markov processes in the dynamic model. Unfortunately, it is not easy to make a statistical analysis of the time-varying amplitude and phase since the relationship between these and the in-phase and quadrature components are highly non-linear. Instead, we make a simulation in Sec. 5 which give some insight into this.

3 Problem Formulation

As stated in the introduction, we take a Bayesian approach to performing the interpolation and making inference about the unknown parameters of the dynamic sinusoidal model in (C.3). In the Bayesian approach, these variables are all random variables, and for the model in (C.3) they are all real and given by

$$
\begin{aligned}
&\text{Observations:} && \boldsymbol{y} = \begin{bmatrix} y_1, y_2, \cdots, y_N \end{bmatrix}^T \\
&\text{Latent variables:} && \boldsymbol{S} = \begin{bmatrix} \boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N \end{bmatrix} \\
&\text{Model parameters:} && \boldsymbol{\theta} = \{\boldsymbol{\omega}, \boldsymbol{\rho}, \boldsymbol{\sigma}_{\boldsymbol{v}}^2, \sigma_w^2\} \\
& && \boldsymbol{\omega} = \begin{bmatrix} \omega_1 & \omega_2 & \cdots & \omega_L \end{bmatrix}^T \\
& && \boldsymbol{\rho} = \begin{bmatrix} \rho_1 & \rho_2 & \cdots & \rho_L \end{bmatrix}^T \\
& && \boldsymbol{\sigma}_{\boldsymbol{v}}^2 = \begin{bmatrix} \sigma_{\boldsymbol{v},1}^2 & \sigma_{\boldsymbol{v},2}^2 & \cdots & \sigma_{\boldsymbol{v},L}^2 \end{bmatrix}^T
\end{aligned}
$$

where $\boldsymbol{s}_n = \begin{bmatrix} \boldsymbol{s}_{n,1}^T, \cdots, \boldsymbol{s}_{n,L}^T \end{bmatrix}^T$ consists of L two-dimensional state vectors pertaining to the L sinusoids. The evolution of these L two-dimensional state vectors is given by (C.8). We also assume that R of the elements in $\boldsymbol{y}$ are missing or corrupted, and that we know their indices $\mathcal{I} \subset \{1, \ldots, N\}$. Using this set of indices, we define the vectors $\boldsymbol{y}_{\mathrm{m}} \triangleq \boldsymbol{y}_{\mathcal{I}}$ and $\boldsymbol{y}_{\mathrm{o}} \triangleq \boldsymbol{y}_{\backslash \mathcal{I}}$ containing the R missing or corrupted observations and the $N - R$ valid observations, respectively. The notation $(\cdot)_{\backslash *}$ denotes 'without element $*$'.

3.1 Inference Aims

The primary aim is to perform the interpolation of the missing or corrupted samples, i.e., to reconstruct the elements of $\boldsymbol{y}_{\mathrm{m}}$ given the valid observations in $\boldsymbol{y}_{\mathrm{o}}$. In classical statistics, this interpolation task is often solved by using an EM-algorithm which iteratively maximises the likelihood function $p(\boldsymbol{y}_{\mathrm{o}}|\boldsymbol{y}_{\mathrm{m}})$, whereas the MAP or MMSE estimate of the posterior distribution $p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}})$ is often used in Bayesian statistics. For the purpose of interpolating music and speech, however, both of these methods tend to produce over-smoothed interpolants in the sense that they do not agree with the

stochastic part of the valid observations [6, 44, 45]. In a Bayesian framework, a much more typical interpolant can be obtained by simply drawing a sample from the posterior distribution $p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}})$.

3.2 Bayesian Inference

The posterior distribution for the missing samples given the valid samples is given by

$$p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}}) = \int p(\boldsymbol{y}_{\mathrm{m}}, \boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}}) d\boldsymbol{S} d\boldsymbol{\theta} \tag{C.18}$$

Unfortunately, we are not able to draw a sample directly from $p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}})$ since we are not able to integrate the nuisance parameters $\boldsymbol{S}$ and $\boldsymbol{\theta}$ out analytically. However, we can obtain a sample from $p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}})$ by taking a single sample from the joint posterior distribution $p(\boldsymbol{y}_{\mathrm{m}}, \boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}})$ and simply ignore the generated values for $\boldsymbol{S}$ and $\boldsymbol{\theta}$. The joint posterior distribution can be written as

$$p(\boldsymbol{y}_{\mathrm{m}}, \boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}}) = p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{S}, \boldsymbol{\theta}, \boldsymbol{y}_{\mathrm{o}}) p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}}) \tag{C.19}$$

where $p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{S}, \boldsymbol{\theta}, \boldsymbol{y}_{\mathrm{o}})$ is known from the observation equation of (C.3). Thus, in order to generate a sample for $\boldsymbol{y}_{\mathrm{m}}$ the only problem left is computing $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}})$. By Bayes' theorem we may write it as

$$p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}}) = \frac{p(\boldsymbol{y}_{\mathrm{o}}, \boldsymbol{S}_{\setminus 1}|\boldsymbol{s}_1, \boldsymbol{\theta}) p(\boldsymbol{s}_1, \boldsymbol{\theta})}{p(\boldsymbol{y}_{\mathrm{o}})} \tag{C.20}$$

where $p(\boldsymbol{y}_{\mathrm{o}}, \boldsymbol{S}_{\setminus 1}|\boldsymbol{s}_1, \boldsymbol{\theta})$, $p(\boldsymbol{s}_1, \boldsymbol{\theta})$ and $p(\boldsymbol{y}_{\mathrm{o}})$ are referred to as the likelihood, the prior and the model evidence, respectively. The likelihood can be factored as

$$p(\boldsymbol{y}_{\mathrm{o}}, \boldsymbol{S}_{\setminus 1}|\boldsymbol{s}_1, \boldsymbol{\theta}) = p(\boldsymbol{y}_{\mathrm{o}}|\boldsymbol{S}_{\setminus \mathcal{I}}, \boldsymbol{\theta}) \prod_{n=1}^{N-1} p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{\theta}) \tag{C.21}$$

which from (C.3) is seen to be a product of normal distributions. Since the state equation of (C.3) decouples into L independent state equations as in (C.8), we can factor the normal distribution $p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{\theta})$ into L bivariate normal distributions given by

$$p(\boldsymbol{s}_{n+1}|\boldsymbol{s}_n, \boldsymbol{\theta}) = \prod_{l=1}^{L} p(\boldsymbol{s}_{n+1,l}|\boldsymbol{s}_{n,l}, \sigma^2_{\boldsymbol{v},l}, \omega_l, \rho_l) \ . \tag{C.22}$$

The form of the prior is considered in Section 3.4. Implicit in the formulation of (C.20) is the model assumption which we consider as known (including its order L)[3]. The

[3] This assumption is quite common although not very realistic.

model evidence in the denominator of (C.20) acts therefore as a mere scale factor since it is independent of $\boldsymbol{S}$ and $\boldsymbol{\theta}$. To reflect this, we simply write Bayes' theorem as

$$p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}}) \propto p(\boldsymbol{y}_{\mathrm{o}}, \boldsymbol{S}_{\setminus 1}|\boldsymbol{s}_1, \boldsymbol{\theta}) p(\boldsymbol{s}_1, \boldsymbol{\theta}) \tag{C.23}$$

where $\propto$ denotes 'proportional to'.

The joint posterior distribution encapsulates all knowledge about the states and model parameters by combining the prior knowledge with the information in the observed data through Bayes' theorem. Theoretically, it is also possible to derive posterior distributions, moments, probability intervals and other posterior characteristics for the individual variables by use of marginalisation and various transformations. In practice, however, it is often either infeasible or impossible to compute these posterior characteristics, and we have to rely on numerical inference methods. The stochastic numerical methods offer various ways of generating samples from the posterior distribution. These samples are then used to form histograms which converge to the true posterior distributions for an increasing sample size. For an overview of some of the methods see, e.g., [22, 23, 46].

3.3 Markov Chain Monto Carlo Sampling

Markov chain Monto Carlo (MCMC) methods are currently a very popular class of stochastic sampling methods adopted by the Bayesian community in the late 1980s [47]. They work by selecting the transition kernel of an ergodic Markov chain such that the invariant distribution of the Markov chain is the desired posterior distribution which we wish to draw samples from. After an initial transient period in which the Markov chain converges, samples generated by the Markov chain are distributed according to the desired distribution. The two most well-known MCMC sampling schemes are the Metropolis-Hastings (MH) algorithm [48–50] and the Gibbs sampler [51]. In the MH algorithm, samples generated from the desired posterior distribution, say $p(\boldsymbol{x})$, which we know up to some normalising constant Z with $p(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x})/Z$, are generated by use of a user-defined proposal distribution $q(\boldsymbol{x}|\boldsymbol{x}^{[\tau]})$ where $\boldsymbol{x}^{[\tau]}$ is the current state of the Markov chain. By construction, $q(\boldsymbol{x}|\boldsymbol{x}^{[\tau]})$ is selected as a trade-off between how similar it is to $p(\boldsymbol{x})$ and how easy it is to generate samples from. A candidate sample $\boldsymbol{x}' \sim q(\boldsymbol{x}|\boldsymbol{x}^{[\tau]})$ is accepted as the next state $\boldsymbol{x}^{[k+1]}$ with probability

$$\alpha(\boldsymbol{x}^{[\tau]}, \boldsymbol{x}') = \min\left[1, \frac{\tilde{p}(\boldsymbol{x}')q(\boldsymbol{x}^{[\tau]}|\boldsymbol{x}')}{\tilde{p}(\boldsymbol{x}^{[\tau]})q(\boldsymbol{x}'|\boldsymbol{x}^{[\tau]})}\right] . \tag{C.24}$$

Otherwise, the current state of the Markov chain is retained. The Gibbs sampler is a special case of the MH-algorithm in which sampling from the multivariate distribution $p(\boldsymbol{x}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)$ is broken up into alternating sampling from the K lower dimensional conditional distribution $p(\boldsymbol{x}_k|\boldsymbol{x}_{\setminus k})$. Specifically, for the k's iteration, we sample

for $k = 1, \ldots, K$ from

$$\boldsymbol{x}_k^{[k+1]} \sim p(\boldsymbol{x}_k | \boldsymbol{x}_1^{[k+1]}, \ldots, \boldsymbol{x}_{k-1}^{[k+1]}, \boldsymbol{x}_{k+1}^{[\tau]}, \ldots, \boldsymbol{x}_K^{[\tau]}) \ . \tag{C.25}$$

The generated samples from these conditional distributions are always accepted.

3.4 Prior Distributions

To complete the Bayesian setup, we need to specify prior distributions on the initial state as well as on the model parameters. In this paper, we assume that we have only vague prior information about the parameters whose joint prior distribution factor as

$$\begin{aligned} p(\boldsymbol{s}_1, \boldsymbol{\theta}) &= p(\boldsymbol{s}_1) p(\boldsymbol{\omega}) p(\boldsymbol{\rho}) p(\boldsymbol{\sigma}_{\boldsymbol{v}}^2) p(\sigma_w^2) \\ &= \left[\prod_{l=1}^{L} p(\boldsymbol{s}_{1,l}) p(\omega_l) p(\rho_l) p(\sigma_{\boldsymbol{v},l}^2) \right] p(\sigma_w^2) \ . \end{aligned} \tag{C.26}$$

For the joint distribution of the l'th frequency parameter and damping coefficient, we use the Jeffreys' prior for the likelihood in (C.22), i.e., $p(\omega_l, \rho_l) = p(\omega_l)p(\rho_l) \propto \rho_l$ for $\omega_l \in [0, \pi]$ and $\rho_l > 0$. It is common to restrict the damping coefficient to be smaller than one since this ensures that the evolution of the in-phase and quadrature components in (C.14) is stable. This yields a beta prior distribution with parameters 2 and 1 on the damping coefficient. In this paper, however, we do not impose this restriction since we wish to model non-stationary signal segments. The selected prior on the frequency parameters causes symmetry in the likelihood of the model parameters which leads to the problem of label switching [52]. More precisely, the joint posterior distribution is a mixture distribution of $L!$ similar distributions up to a permutation of labels [28]. For the interpolation of missing samples, which is the primary focus of this paper, this is not a problem. For making inference about the unknown parameters, however, the problem can be addressed by ensuring identifiability of the frequency parameters through a joint prior distribution on the frequency parameters given by

$$p(\boldsymbol{\omega}) \propto \mathbb{I}_{[0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_L \leq \pi]}(\boldsymbol{\omega}) \tag{C.27}$$

where $\mathbb{I}_{[A]}(\cdot)$ is the indicator function on the region A. Alternatively, the generated samples can also be postprocessed by applying various clustering techniques to the generated frequency parameters [52].

For the observation and state noise variances, we use inverse gamma distributions, i.e., $p(\sigma_w^2) = \text{Inv-}\mathcal{G}(\alpha_w, \beta_w)$ and $p(\sigma_{\boldsymbol{v},l}^2) = \text{Inv-}\mathcal{G}(\alpha_{v,l}, \beta_{v,l})$. These distributions can be made diffuse by choosing small values for the hyperparameters. They can also be used for preventing the noise variances from collapsing to zero which is often a necessary requirement in MCMC based inference [5]. For the initial state distribution, we assume a normal distribution, i.e., $p(\boldsymbol{s}_1) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{P})$.

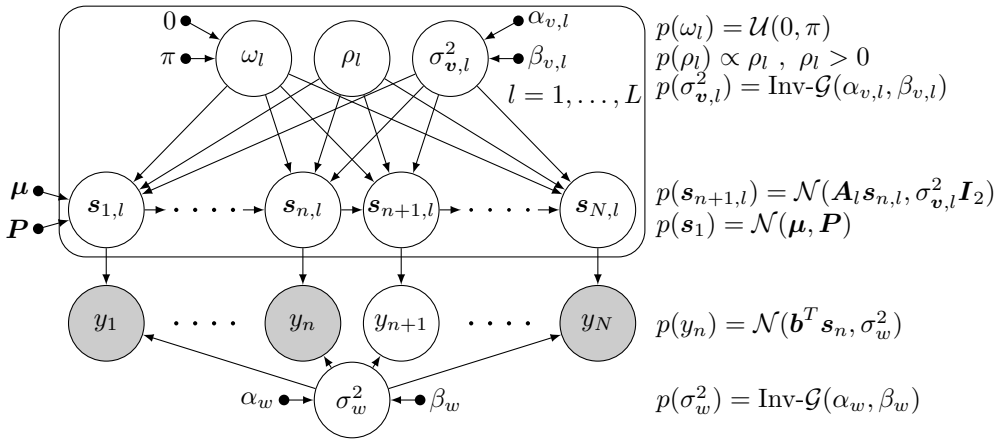

Fig. C.1: A directed graphical model for the parameter estimation and interpolation problem. Shaded nodes denote observed variables (notice that y_{n+1} is a missing observation), open circles denote latent variables, and smaller solid circles denote deterministic hyperparameters. The box denotes a plate, and it is labelled with $l = 1, \ldots, L$ indicating that there are L copies of the nodes inside.

4 Derivation of Inference Scheme

The Bayesian model considered in the previous section is summarised in the directed graphical model in Figure C.1. The figure clearly reveals the assumptions, the conditional dependency between the variables, and the hierarchical structure to the setup also given by likelihood in (C.21) and (C.22), and the prior in (C.26). In our inference scheme for the variables of the model, we draw samples from the joint posterior distribution $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\text{o}})$ by means of a Gibbs sampler. As detailed in Section 3.3, we therefore have to group the variables into suitable blocks and derive conditional distributions for them. In this paper, we consider the following two conditional distributions given by

$$\text{States:} \qquad p(\boldsymbol{S}|\boldsymbol{\theta}, \boldsymbol{y}_{\text{o}}) \tag{C.28}$$

$$\text{Model parameters:} \qquad p(\boldsymbol{\theta}|\boldsymbol{S}, \boldsymbol{y}_{\text{o}}) \tag{C.29}$$

The selected grouping of variables in (C.28) and (C.29) leads to a set of conditional distributions which are fairly easy to sample from. Further, by sampling all model parameters in a single step, we increase the mixing properties of the sampler, i.e., we decrease the correlation of the generated samples leading to faster convergence of the underlying Markov chain. In the next section, we derive the particular form of these conditional distributions.

4.1 States

The conditional state distribution in (C.28) is a multivariate Gaussian distribution. However, the dimension of a sample from this distribution is $2LN \times 1$ which would render direct sampling from it infeasible for most applications. Instead, we use the simulation smoother for drawing samples from (C.28). The simulation smoother is an efficient sampling scheme using standard Kalman smoothing, and it is easily modified to handle the case of missing observations since this corresponds to skipping the update step of the build-in Kalman filter for these samples. The simulation smoother exists in several versions of which we use the version in [53] (see, e.g., [54–56] for other versions of the simulation smoother).

4.2 Model Parameters

Since the model parameter σ_w^2 of the observation equation and the L sets of model parameters $(\omega_l, \rho_l, \sigma_{\boldsymbol{v},l}^2)$ of the state equation are mutually independent conditioned on the states $\boldsymbol{S}$, we can factor (C.29) as

$$p(\boldsymbol{\theta}|\boldsymbol{S}, \boldsymbol{y}_{\mathrm{o}}) = \left[\prod_{l=1}^{L} p(\omega_l, \rho_l, \sigma_{\boldsymbol{v},l}^2|\boldsymbol{S})\right] p(\sigma_w^2|\boldsymbol{S}, \boldsymbol{y}_{\mathrm{o}}) \ . \tag{C.30}$$

Thus, sampling from the conditional distribution in (C.29) can be done by sampling the $L+1$ conditional distributions on the right hand side of (C.30) independently.

Frequency, Damping and State Noise Variance

The main difficulty of our Gibbs sampler is to draw samples from the joint conditional distribution of the frequency parameter, the damping coefficient, and the state noise variance given the states, i.e., $p(\omega_l, \rho_l, \sigma_{\boldsymbol{v},l}^2|\boldsymbol{S})$. To our knowledge, it is not possible to sample directly from $p(\omega_l, \rho_l, \sigma_{\boldsymbol{v},l}^2|\boldsymbol{S})$ - although we come close in this paper. A Gibbs sampling scheme on the individual parameters is also not straight-forward since it suffers from poor mixing and since the l'th damping coefficient conditioned on the l'th frequency parameter and state noise variance has a non-standard distribution. In order to improve mixing, we therefore propose sampling all parameters at once from $p(\omega_l, \rho_l, \sigma_{\boldsymbol{v},l}^2|\boldsymbol{S})$ by use of an MH-sampler previously discussed in Section 3.3. For the proposed MH-sampler the candidate samples are easy to generate and the acceptance probability turns out to be very easy to evaluate.

Given the states, the posterior distribution for the l'th set of model parameters of

the state equation can be written as

$$p(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S}) \propto \left[\prod_{n=1}^{N-1} p(\boldsymbol{s}_{n+1,l}|\boldsymbol{s}_{n,l}, \sigma^2_{\boldsymbol{v},l}, \omega_l, \rho_l)\right] \times p(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}) \tag{C.31}$$

where $p(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l})$ is the prior distribution which, as stated in Section 3.4, factors into

$$p(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}) = p(\omega_l)p(\rho_l)p(\sigma^2_{\boldsymbol{v},l})$$

The distribution for $p(\boldsymbol{s}_{n+1,l}|\boldsymbol{s}_{n,l}, \sigma^2_{\boldsymbol{v},l}, \omega_l, \rho_l)$ is a bivariate normal distribution and the product over n of $N-1$ of these can therefore be written as

$$\prod_{n=1}^{N-1} p(\boldsymbol{s}_{n+1,l}|\boldsymbol{s}_{n,l}, \sigma^2_{\boldsymbol{v},l}, \omega_l, \rho_l) \propto {\sigma^2_{\boldsymbol{v},l}}^{-(N-1)} \times \exp\left\{\frac{-1}{2\sigma^2_{\boldsymbol{v},l}} \sum_{n=1}^{N-1} (\boldsymbol{s}_{n+1,l} - \boldsymbol{A}_l\boldsymbol{s}_{n,l})^T(\boldsymbol{s}_{n+1,l} - \boldsymbol{A}_l\boldsymbol{s}_{n,l})\right\} . \tag{C.32}$$

In order to write this distribution in a useful way in terms of the frequency parameter and the damping coefficient, we rewrite $\boldsymbol{A}_l\boldsymbol{s}_{n,l}$ into

$$\boldsymbol{A}_l\boldsymbol{s}_{n,l} = \begin{bmatrix}\boldsymbol{s}_{n,l} & \boldsymbol{s}^{\perp}_{n,l}\end{bmatrix} \boldsymbol{a}_l \tag{C.33}$$

where $\boldsymbol{s}^{\perp}_{n,l}$ is obtained by a 90° clockwise rotation of $\boldsymbol{s}_{n,l}$ and

$$\boldsymbol{a}_l \triangleq \rho_l \begin{bmatrix}\cos\omega_l & \sin\omega_l\end{bmatrix}^T . \tag{C.34}$$

Inserting this into (C.32) and replacing the summation with an inner product yield

$$\prod_{n=1}^{N-1} p(\boldsymbol{s}_{n+1,l}|\boldsymbol{s}_{n,l}, \sigma^2_{\boldsymbol{v},l}, \omega_l, \rho_l) = \mathcal{N}(\boldsymbol{\varphi}_l; \boldsymbol{\Phi}_l\boldsymbol{a}_l, \sigma^2_{\boldsymbol{v},l}\boldsymbol{I}_{2(N-1)})$$

where we have defined

$$\boldsymbol{\varphi}_l \triangleq \begin{bmatrix}\boldsymbol{s}^T_{2,l} & \boldsymbol{s}^T_{3,l} & \cdots & \boldsymbol{s}^T_{N,l}\end{bmatrix}^T \tag{C.35}$$

$$\boldsymbol{\phi}_l \triangleq \begin{bmatrix}\boldsymbol{s}^T_{1,l} & \boldsymbol{s}^T_{2,l} & \cdots & \boldsymbol{s}^T_{N-1,l}\end{bmatrix}^T \tag{C.36}$$

$$\tilde{\boldsymbol{\phi}}_l \triangleq \begin{bmatrix}(\boldsymbol{s}^{\perp}_{1,l})^T & (\boldsymbol{s}^{\perp}_{2,l})^T & \cdots & (\boldsymbol{s}^{\perp}_{N-1,l})^T\end{bmatrix}^T \tag{C.37}$$

$$\boldsymbol{\Phi}_l \triangleq \begin{bmatrix}\boldsymbol{\phi}_l & \tilde{\boldsymbol{\phi}}_l\end{bmatrix} . \tag{C.38}$$

Now, by assuming a non-informative prior for $\boldsymbol{a}_l$ of the form

$$p(\boldsymbol{a}_l|\sigma^2_{\boldsymbol{v},l}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{v},l}\delta\boldsymbol{I}_2) \qquad \text{with } \delta \to \infty \tag{C.39}$$

and by using standard Bayesian inference for the linear model [21], we obtain after some algebra

$$p(\boldsymbol{a}_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S}) \propto p(\boldsymbol{S}|\boldsymbol{a}_l, \sigma^2_{\boldsymbol{v},l})p(\boldsymbol{a}_l|\sigma^2_{\boldsymbol{v},l})p(\sigma^2_{\boldsymbol{v},l}) \tag{C.40}$$

$$\propto \mathcal{NIG}(\boldsymbol{\mu}_{\boldsymbol{a},l}, \sigma^2_{\boldsymbol{a},l}\boldsymbol{I}_2, \alpha_{q,l}, \beta_{q,l}) \ . \tag{C.41}$$

where the parameters of the normal-scaled inverse gamma distribution are defined by

$$\sigma^2_{\boldsymbol{a},l} \triangleq (\boldsymbol{\phi}_l^T\boldsymbol{\phi}_l)^{-1} \tag{C.42}$$

$$\boldsymbol{\mu}_{\boldsymbol{a},l} \triangleq \sigma^2_{\boldsymbol{a},l}\boldsymbol{\Phi}_l^T\boldsymbol{\varphi}_l \tag{C.43}$$

$$\alpha_{\sigma^2_{\boldsymbol{v},l}} \triangleq \alpha_{v,l} + N - 1 \tag{C.44}$$

$$\beta_{\sigma^2_{\boldsymbol{v},l}} \triangleq \beta_{v,l} + (\boldsymbol{\varphi}_l^T\boldsymbol{\varphi}_l - \sigma^{-2}_{\boldsymbol{a},l}\boldsymbol{\mu}^T_{\boldsymbol{a},l}\boldsymbol{\mu}_{\boldsymbol{a},l})/2 \ . \tag{C.45}$$

The Jacobian determinant of the transformation from $\boldsymbol{a}_l$ to (ω_l, ρ_l) is given by

$$\left|\frac{\partial\boldsymbol{a}_l}{\partial\rho_l} \quad \frac{\partial\boldsymbol{a}_l}{\partial\omega_l}\right| = \begin{vmatrix}\cos\omega_l & -\rho_l\sin\omega_l \\ \sin\omega_l & \rho_l\cos\omega_l\end{vmatrix} = \rho_l \ ,$$

which is proportional to the prior distribution on the damping coefficient. Therefore, we may write (C.41) as

$$q(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S}) \propto p(\rho_l)p(\boldsymbol{a}_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S}) \tag{C.46}$$

with $\boldsymbol{a}_l$ replaced by the expression in (C.34). Thus, the distribution $q(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S})$ is nearly identical to the desired distribution $p(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S})$ in (C.31). The only difference between the two distributions is that the frequency parameter of $q(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S})$ is uniform on $[-\pi, \pi]$ whereas is it uniform on $[0, \pi]$ in $p(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S})$. In order to remedy for this, we use $q(\omega_l, \rho_l, \sigma^2_{\boldsymbol{v},l}|\boldsymbol{S})$ as a proposal distribution in an MH-sampler. We draw a sample from this proposal by first sampling a set $(\boldsymbol{a}_l', {\sigma^2_{\boldsymbol{v},l}}')$ from the bivariate normal-scaled inverse gamma distribution in (C.41). Sampling from the bivariate normal-scaled inverse gamma distribution can be done in various ways. Here, we sample from its marginal densities given by

$$p(\sigma^2_{\boldsymbol{v},l}|\boldsymbol{S}) = \text{Inv-}\mathcal{G}(\alpha_{\sigma^2_{\boldsymbol{v},l}}, \beta_{\sigma^2_{\boldsymbol{v},l}}) \tag{C.47}$$

$$p(\boldsymbol{a}_l|\boldsymbol{S}) = \mathcal{T}\left(\boldsymbol{\mu}_{\boldsymbol{a},l}, \frac{\beta_{\sigma^2_{\boldsymbol{v},l}}}{\alpha_{\sigma^2_{\boldsymbol{v},l}}}\sigma^2_{\boldsymbol{a},l}\boldsymbol{I}_2, 2\alpha_{\sigma^2_{\boldsymbol{v},l}}\right) \ . \tag{C.48}$$

This is done by sampling from [46]

$$\sigma_{\boldsymbol{v},l}^{2}{}' \sim \text{Inv-}\mathcal{G}(\alpha_{\sigma_{\boldsymbol{v},l}^2}, \beta_{\sigma_{\boldsymbol{v},l}^2}) \tag{C.49}$$

$$\tau_l' \sim \text{Inv-}\mathcal{G}(\alpha_{\sigma_{\boldsymbol{v},l}^2}, 1/2) \tag{C.50}$$

$$\boldsymbol{a}_l' = \begin{bmatrix} a_{1,l}' & a_{2,l}' \end{bmatrix}^T \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{a},l}, 2\beta_{\sigma_{\boldsymbol{v},l}^2}\tau_l'\sigma_{\boldsymbol{a},l}^2\boldsymbol{I}_2) \ . \tag{C.51}$$

Second, we transforms the generated sample $(\boldsymbol{a}_l', \sigma_{\boldsymbol{v},l}^{2}{}')$ into $(\omega_l', \rho_l', \sigma_{\boldsymbol{v},l}^{2}{}')$ by use of the transformation

$$\omega_l' = \arctan(a_{2,l}'/a_{1,l}') \tag{C.52}$$

$$\rho_l' = \sqrt{\boldsymbol{a}_l'^T\boldsymbol{a}_l'} \ . \tag{C.53}$$

Finally, the samples generated by this proposal distribution are accepted with probability

$$\begin{aligned}
&\alpha\big((\omega_l, \rho_l, \sigma_{\boldsymbol{v},l}^2), (\omega_l', \rho_l', \sigma_{\boldsymbol{v},l}^{2}{}')\big) \\
&= \min\left[1, \frac{p(\boldsymbol{a}_l', \sigma_{\boldsymbol{v},l}^{2}{}'|\boldsymbol{S})p(\omega_l')p(\rho_l')q(\omega_l, \rho_l, \sigma_{\boldsymbol{v},l}^2|\boldsymbol{S})}{p(\boldsymbol{a}_l, q_l|\boldsymbol{S})p(\omega_l)p(\rho_l)q(\omega_l', \rho_l', \sigma_{\boldsymbol{v},l}^{2}{}'|\boldsymbol{S})}\right] \\
&= \min\left[1, \frac{p(\omega_l')}{p(\omega_l)}\right] \\
&= \mathbb{I}_{[0,\pi]}(\omega_l') \ .
\end{aligned} \tag{C.54}$$

If the sample is not accepted, the previous values $(\omega_l, \rho_l, \sigma_{\boldsymbol{v},l}^2)$ are retained. In the case where we use the structured prior on the frequency parameters, the indicator function should be changed to $\mathbb{I}_{[\omega_{l-1},\omega_{l+1}]}(\omega_l')$.

Observation Noise Variance

By Bayes' theorem, we can write $p(\sigma_w^2|\boldsymbol{S}, \boldsymbol{y}_\text{o})$ as

$$p(\sigma_w^2|\boldsymbol{S}, \boldsymbol{y}_\text{o}) \propto p(\boldsymbol{y}_\text{o}|\boldsymbol{S}_{\setminus\mathcal{I}}, \sigma_w^2)p(\sigma_w^2) \tag{C.55}$$

where $p(\boldsymbol{y}_\text{o}|\boldsymbol{S}_{\setminus\mathcal{I}}, \sigma_w^2)$ is the likelihood of the observation equation in (C.3) and $p(\sigma_w^2)$ is the prior distribution for σ_w^2. Since $p(\boldsymbol{y}_\text{o}|\boldsymbol{S}_{\setminus\mathcal{I}}, \sigma_w^2) = \mathcal{N}(\boldsymbol{S}_{\setminus\mathcal{I}}^T\boldsymbol{b}, \sigma_w^2\boldsymbol{I}_{N-R})$ and $p(\sigma_w^2) = \text{Inv-}\mathcal{G}(\alpha_w, \beta_w)$, the posterior distribution $p(\sigma_w^2|\boldsymbol{S}_{\setminus\mathcal{I}}, \boldsymbol{y}_\text{o})$ is an inverse gamma distribution, $\text{Inv-}\mathcal{G}(\sigma_w^2; \alpha_{\sigma_w^2}, \beta_{\sigma_w^2})$, with parameters

$$\alpha_{\sigma_w^2} = \alpha_w + N/2 \tag{C.56}$$

$$\beta_{\sigma_w^2} = \beta_w + \frac{1}{2}(\boldsymbol{y}_\text{o} - \boldsymbol{S}_{\setminus\mathcal{I}}^T\boldsymbol{b})^T(\boldsymbol{y}_\text{o} - \boldsymbol{S}_{\setminus\mathcal{I}}^T\boldsymbol{b}) \ . \tag{C.57}$$

Table C.1: Summary of proposed Gibbs sampler for generating samples from $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\text{o}})$.

1. Select hyperparameters and initialise the Gibbs sampler.
2. Repeat for $k = 0, 1, 2, \ldots, K$
   (a) $\boldsymbol{S}^{[k+1]} \sim p(\boldsymbol{S}|\boldsymbol{\theta}^{[k]}, \boldsymbol{y}_{\text{o}})$ (simulation smoother)
   (b) Repeat for $l = 1, 2, \ldots, L$
      i. $\sigma_{\boldsymbol{v},l}^{2\,\prime} \sim \text{Inv-}\mathcal{G}(\alpha_{\sigma_{\boldsymbol{v},l}^2}^{[\tau]}, \beta_{\sigma_{\boldsymbol{v},l}^2}^{[\tau]})$
      ii. $\tau_l' \sim \text{Inv-}\mathcal{G}(\alpha_{\sigma_{\boldsymbol{v},l}^2}^{[\tau]}, 1/2)$
      iii. $\boldsymbol{a}_l' \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{a},l}^{[\tau]}, 2\beta_{\sigma_{\boldsymbol{v},l}^2}^{[\tau]} \tau_l' \sigma_{\boldsymbol{a},l}^{2\,[\tau]} \boldsymbol{I}_2)$
      iv. $\omega_l' = \arctan(a_{2,l}'/a_{1,l}')$
      v. $\rho_l' = \sqrt{\boldsymbol{a}_l'^T \boldsymbol{a}_l'}$
      vi. if $\omega_l' > 0$
         - $(\omega_l^{[k+1]}, \rho_l^{[k+1]}, \sigma_{\boldsymbol{v},l}^{2\,[k+1]}) = (\omega_l', \rho_l', \sigma_{\boldsymbol{v},l}^{2\,\prime})$

      else
         - $(\omega_l^{[k+1]}, \rho_l^{[k+1]}, \sigma_{\boldsymbol{v},l}^{2\,[k+1]}) = (\omega_l^{[\tau]}, \rho_l^{[\tau]}, \sigma_{\boldsymbol{v},l}^{2\,[\tau]})$

   (c) $\sigma_w^{2\,[k+1]} \sim \text{Inv-}\mathcal{G}(\alpha_{\sigma_w^2}, \beta_{\sigma_w^2}^{[\tau]})$

4.3 Summary of Inference Scheme

Table C.1 summarises our proposed Gibbs sampler for generating samples from $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\text{o}})$. The computational complexity of the algorithm is fairly high primarily due to the generation of the states by the simulation smoother. In our implementation with $N = 600$ observations and $L = 6$ sinusoids, it takes approximately 40 ms for generating a state sample $\boldsymbol{S}^{[\tau]}$. This corresponds to nearly 97 % of the time consumption of one iteration of the Gibbs sampler. For the application of interpolation, we only need a single sample for the states and model parameters from the invariant distribution of the underlying Markov chain of the sampler. Once these have been generated, we may perform the interpolation by simulating from the observation equation of (C.3). Therefore, the computational complexity of the algorithm heavily depends on proper initialisation and the convergence speed of the chain.

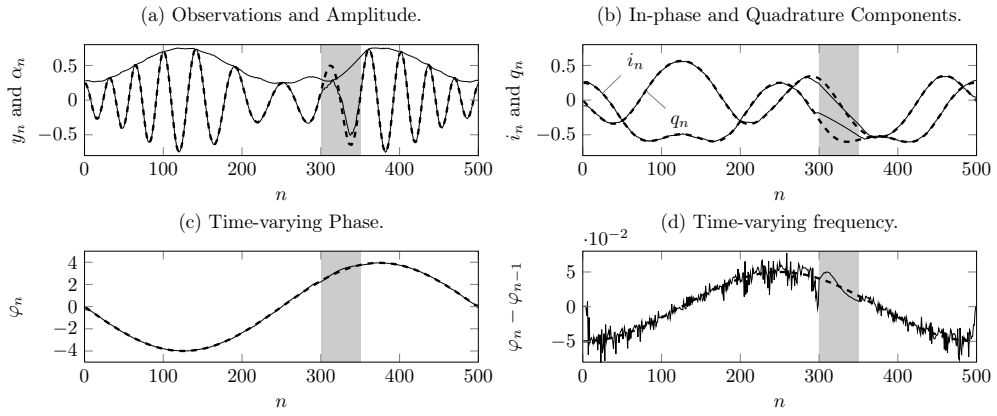

Fig. C.2: Synthetic signal with both amplitude and phase modulation. The shaded area indicates a missing section of the signal. Plot (a) shows the signal (dashed line), a state generated by our Gibbs sampler, and the average amplitude. Plot (b) shows the true (dashed lines) and the average in-phase and quadrature components. Plot (c) and (d) show the true (dashed lines) and the average phase and time-varying part of the frequency.

5 Simulations

In this section, we first demonstrate that the dynamic signal model is able to model signals with amplitude and frequency modulation. These phenomena are encountered in real world signals. Second, we illustrate the proposed inference scheme on a synthetic signal and apply it to the application of reconstructing missing or corrupted audio packets on a packet-based network[4]. In our simulations, we use the following common setup for our Gibbs sampler. We use non-informative prior distributions with hyperparameters

$$\begin{bmatrix} \boldsymbol{\mu} & \boldsymbol{P} & \alpha_i & \beta_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} & 10\boldsymbol{I}_2 & 0 & 10^{-5} \end{bmatrix}.$$

where $i = \{v, w\}$. The small non-zero value for β_i is selected in order to prevent the noise variances from collapsing to zero. The Gibbs sampler is iterated $10,000$ times and samples from the first $1,000$ iterations are discarded as burn-in samples. The initial values for the frequency and observation noise variance are found by using a simple matching pursuit algorithm [57]. The initial value for the damping coefficient and state noise variance are set to 1 and $\sigma_w^{2\,[0]}/10$, respectively. For the model order, we use $L = 1$ in the two examples with synthetic signals and $L = 6$ in the two examples with real world signals.

[4]The MATLAB code and audio samples used in the simulations can be obtained from http://kom.aau.dk/~jkn/publications/publications.php

5.1 Applicability of the Model

The static model in (C.2) is very useful for modelling the periodic parts of a signal. However, since the phase and frequency are modelled as constants and the amplitude with an exponentially decaying envelope, the static model is in general not able to capture common phenomena such as amplitude and frequency modulation [10]. As discussed in Sec. 2, the dynamic model allows the in-phase and quadrature components to develop as a first order Gauss-Markov process. Thus, the model also allows the amplitude, the phase and hence the frequency to be time-varying. These are given by

$$\alpha_{n,l} = \rho_l^n \sqrt{i_{n,l}^2 + q_{n,l}^2} \tag{C.58}$$

$$\varphi_{n,l} = \arctan(q_{n,l}/i_{n,l}) \tag{C.59}$$

$$\omega_{n,l} = \omega_l + d\varphi_{t,l}/dt\big|_{t=nT} \tag{C.60}$$

where the time-varying frequency $\omega_{n,l}$ is a sum of the frequency ω_l from the dynamic model in (C.3) and the sampled derivative of the continuous-time phase $\varphi_{t,l}$.

In Fig. C.2.a, we have shown a synthetic signal consisting of a single sinusoid with both sinusoidal amplitude and frequency modulation (dashed line). The signal consists of $N = 500$ samples and is given by

$$x_n = \alpha_n \cos(\theta_n) + w_n \tag{C.61}$$

$$\alpha_n = 0.5 + 0.25 \sin(4\pi n/N - \pi/2) \tag{C.62}$$

$$\theta_n = 0.15n - 0.05 \sum_{m=1}^{n} \sin(2\pi m/N - \pi/2) \tag{C.63}$$

where w_n is white Gaussian noise with variance 10^{-6}. The samples from index 300 to index 350 were removed and considered to be missing samples. We used the proposed inference scheme for analysing the signal x_n, and the full line on top of the dashed line in Fig. C.2.a shows a state vector generated by our Gibbs sampler. For all generated state samples, we also demodulated the states in order to obtain the samples for the in-phase and quadrature components. Based on these samples, we calculated the average amplitude $\alpha_{n,l}$, the average in-phase and quadrature components, the average phase $\varphi_{n,l}$ and the average derivative of the phase as $\varphi_{n,l} - \varphi_{n-1,l}$. The latter is an approximation to the derivative of the phase. These averages (full lines) are compared against their true values (dashed lines) in Fig. C.2.a, Fig. C.2.b, Fig. C.2.c and Fig. C.2.d, respectively. Clearly, the model is able to capture both amplitude and frequency modulation. However, the figures also reveal a potential problem for the application of interpolating missing samples; In this example, the in-phase and quadrature components do not evolve as a typical Gauss-Markov process. Therefore, we cannot expect the interpolation to be very successful since our interpolation scheme, on average, will reconstruct the missing samples in the in-phase and quadrature components with a straight line.

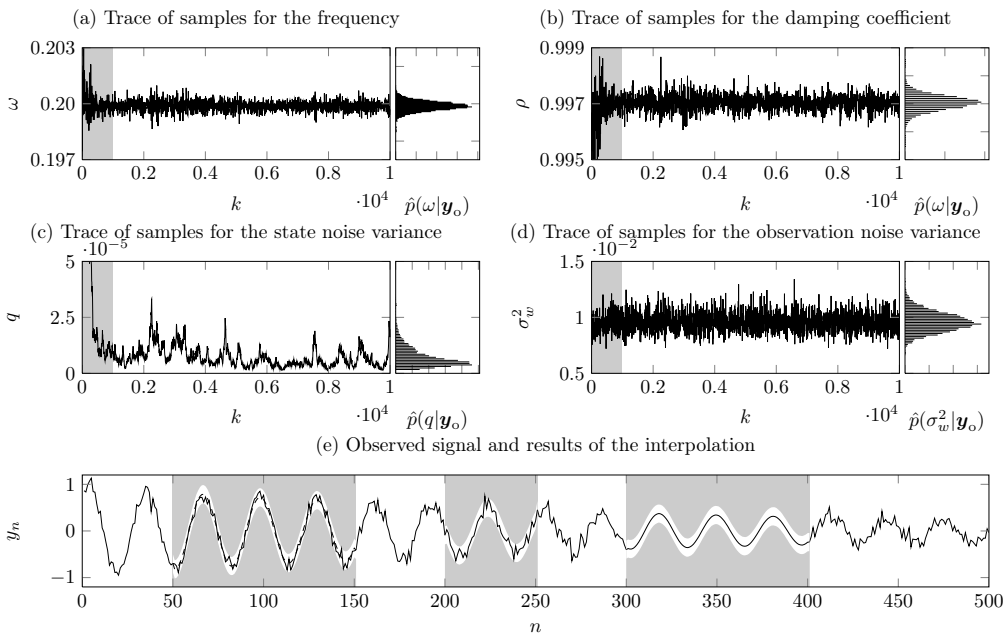

Fig. C.3: Traces of 10,000 generated samples for the (a) frequency, (b) damping coefficient, (c) state noise variance, and (d) observation noise variance. The histograms in the right margin of plot (a)-(d) are computed based on the last 9,000 samples. Only 50 % of the signal is observed and the three missing sections are indicated by a shaded background. In the three interpolation sections, the 95 % credible interval for the posterior distribution $\hat{p}(\boldsymbol{y}_m|\boldsymbol{y}_o)$ of the interpolated samples are shown along with interpolated samples based on the mean of $\hat{p}(\boldsymbol{y}_m|\boldsymbol{y}_o)$ (last section), a sample from $\hat{p}(\boldsymbol{y}_m|\boldsymbol{y}_o)$ (middle section) and both (first section).

5.2 Synthetic Signal

We consider a simple synthetic signal generated by the static sinusoidal model. We do this in order to illustrate some of the features of the proposed Bayesian inference scheme and interpolator. Specifically, we generated $N = 500$ observations from the static model in (C.1) with a single sinusoidal component with parameters

$$\begin{bmatrix}\alpha & \beta & \omega & \rho & \sigma_w^2\end{bmatrix} = \begin{bmatrix}1/\sqrt{2} & 1/\sqrt{2} & 0.2 & 0.997 & 0.01\end{bmatrix}.$$

We also removed 50 % of the observations distributed over three sections as illustrated in Fig. C.3.e. Fig. C.3 shows the results of running the Gibbs sampler. In Fig. C.3.a-C.3.d, the traces of the 10,000 generated samples for the model parameters are shown, and Fig. C.3.e shows the results of interpolating the sections of missing observations. The underlying Markov chain seems to have converged to its invariant distribution after approximately 500 samples. The histograms in the margin of the first four plots are based on the last 9,000 generated samples. They are an approximation to the marginal distribution for the individual model parameters, and they converge to it for an increasing number of iterations of the Gibbs sampler. As previously discussed in Section 4, the histograms can be used for summarising various posterior features such as point and interval estimates. For example, computing their means yields the estimates $\hat{\omega} = 0.1999$, $\hat{\rho} = 0.997 \cdot 10^{-3}$, $\hat{q} = 7.044 \cdot 10^{-6}$, and $\hat{\sigma}_w^2 = 9.606 \cdot 10^{-3}$.

In Fig. C.3.e, the three interpolation sections are shown with a shaded background. In all three simulation sections, we have shown the 95 % credible interval for the missing observations[5]. In the last interpolation section, we have used the mean estimate of the interpolated samples whereas the interpolation in the middle section is a random sample from the posterior distribution. Both methods are shown in the first interpolation section. Clearly, sampling from the posterior distribution yields a much more typical sample than using the mean estimate. The latter has higher probability, but it does not model the noise.

5.3 Music Signal

In the third simulation, we considered a segment of observations from a downsampled trumpet signal whose spectrogram can be seen in Fig. C.4.b. The considered snapshot corresponds to 75 ms of audio and is shown in Fig. C.4.d. The periodogram of the $N = 660$ observations in the snapshot is shown in Fig. C.4.c. Prior to running the Gibbs sampler, we removed the middle section thus emulating a lost audio packet of 25 ms on a packet-based network. In Fig. C.4.a, we have shown the six traces of samples for the frequencies. We see that the sampler reached a stationary point after approximately

[5]The credible intervals were computed by assuming that the missing observations were normally distributed. More precise, but also more complex, methods for estimating the credible interval can be found in, e.g., [58].

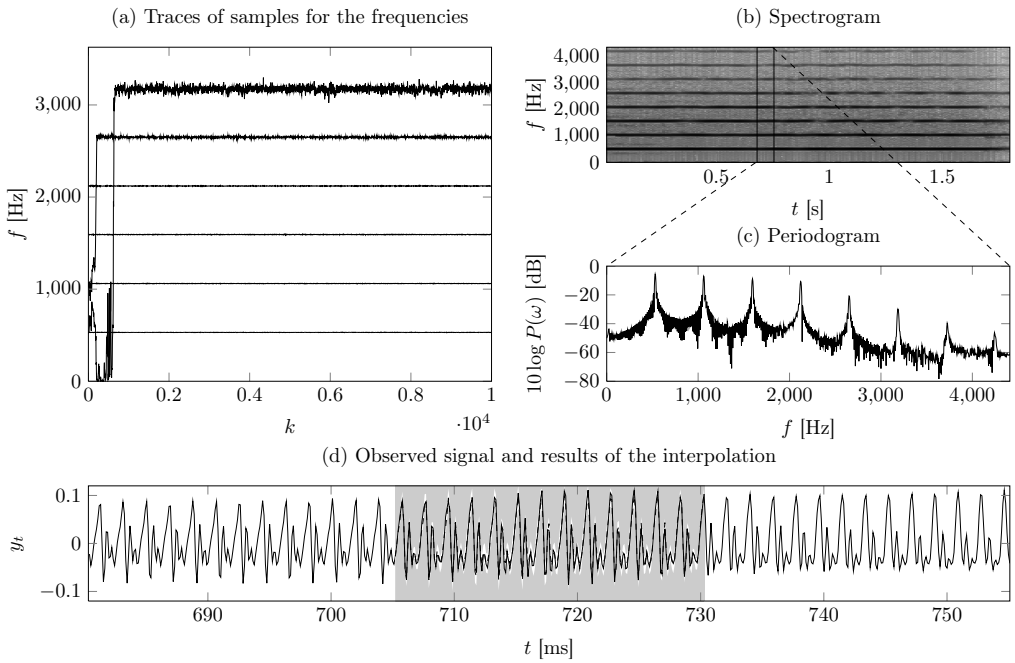

Fig. C.4: Plot (a) shows the six traces for the frequencies each consisting of 10,000 samples. Plot (b) shows the spectrogram for the complete trumpet signal whereas plot (c) shows the periodogram for the section indicated in plot (b). The time series corresponding to this section is shown in plot (d) with the middle section of 25 ms audio missing. The plot also shows the result of the interpolation in terms of 95 % probability interval, a sample for the posterior distribution $\hat{p}(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}})$ and the true missing observations (dotted).

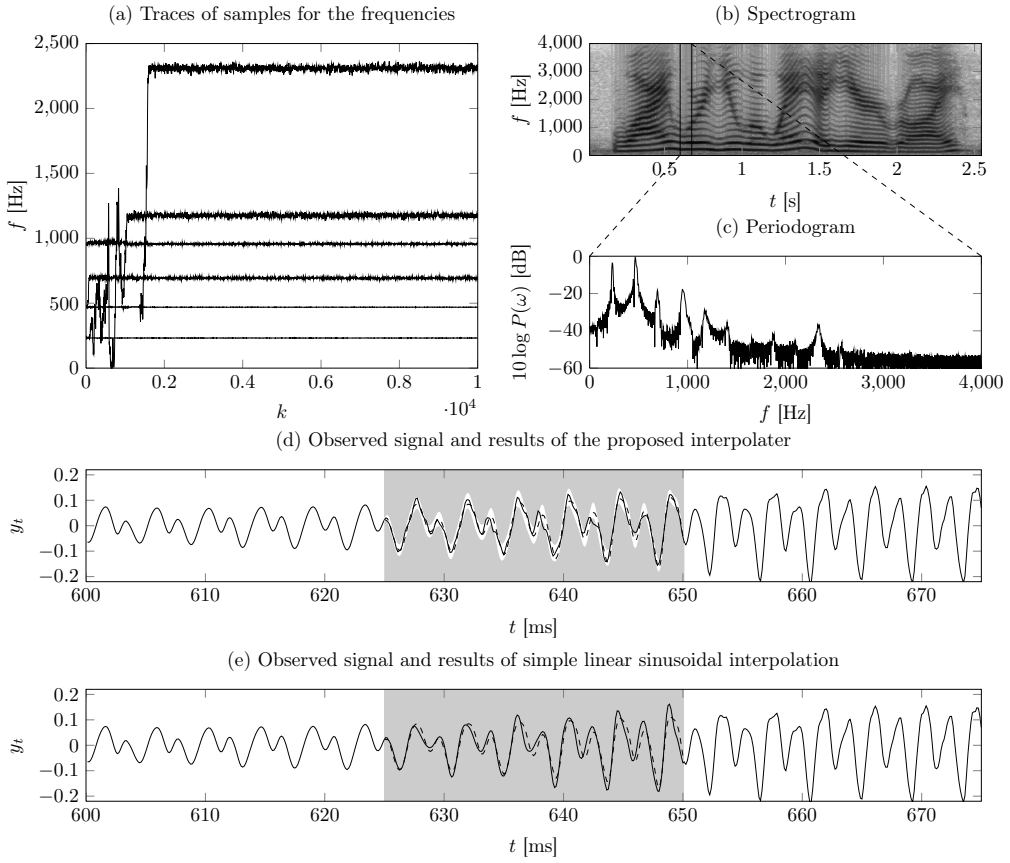

Fig. C.5: Plot (a) shows the six traces for the frequencies each consisting of 10,000 samples. Plot (b) shows the spectrogram for the complete speech signal whereas plot (c) shows the periodogram for the section indicated in plot (b). The time series corresponding to this section is shown in plot (d) with the middle section of 25 ms audio missing. The plot also shows the result of the interpolation in terms of 95 % probability interval, a sample for the posterior distribution $\hat{p}(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}})$ and the true missing observations (dotted). For comparison, the missing packet was also interpolated in plot (e) by use of a simpler interpolator.

500 iterations after which samples for the dominating six frequency components were generated. The results of the interpolation are shown in Fig. C.4.d. It is observed, that the 95 % credible interval was very tight and that the generated sample from the posterior distribution for the missing observations therefore almost coincided with the true missing observations. An informal listening test also confirmed that the music segment had been restored with almost no perceptual loss.

5.4 Speech Signal

In the fourth and final simulation, we considered a more challenging segment of observations originating from a speech signal; where the frequency spectrum and amplitudes of the trumpet signal in the previous simulation were approximately constant, the snapshot shown in Fig. C.5.d is clearly non-stationary. Additionally, as can be seen from the spectrogram in Fig. C.5.b, some of the frequencies are non-constant in the snapshot. The speech signal originates from a female voice uttering *"Why were you away a year, Roy?"*, and it was downsampled to $8,000$ Hz. As in the previous simulation, we removed the middle section of 25 ms prior to running the Gibbs sampler. The traces of samples for the frequencies are shown in Fig. C.5.a. The sampling scheme seemed to have reached a stationary point after approximately 1500 iterations. The interpolated samples in Fig. C.5.d follows the same increasing trend as the true signal. Compared against the interpolation of the trumpet signal, the 95 % confidence interval is wider reflecting the more complex structure of the signal. Despite this, an informal listening test revealed that the music segment had been restored with only little perceptual loss. For comparison, we have also performed the interpolation of the missing packet by use of a simpler interpolater based on [9]. In this interpolation scheme, the amplitudes and frequencies are estimated on both sides of the missing packet which is recovered by linearly interpolating these amplitudes and frequencies. The result of this interpolation is shown in Fig. C.5.e. In order to compare the two methods, we have measured the reconstruction signal-to-noise ratio (SNR) for both methods. For the simple linear sinusoidal interpolater, the SNR was 7.7 dB whereas a sample from the posterior distribution $p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}})$ resulted in an SNR of 10.8 dB. If we instead used the posterior mean as an interpolant, the SNR was 15.8 dB. It should be noted, however, that SNR cannot be used as an objective measure for the reconstruction performance since the human auditory system does not perceive sound degradation in the two norm.

6 Conclusion

In this paper, we have presented a Bayesian interpolation and parameter estimation inference scheme based on a dynamic signal model hypothesis for the observed segment of data. The dynamic model enables modelling of real world signals with non-stationary, but smooth evolution since the in-phase and quadrature components were modelled as first order Gauss-Markov processes. The proposed inference scheme for the dynamic model was developed in a Bayesian framework and comprised two stages. In the first stage, a two state Gibbs sampler alternated between sampling from the conditional distribution for the hidden states given the model parameters and sampling from the conditional distribution for the model parameters given the hidden states. In the second stage, a single draw from the posterior distribution for the missing observations given the last sample for the hidden states and model parameters was obtained. This sample

was used for replacing the missing sample with a typical interpolant for the underlying process.

In the simulations, we demonstrated that the inference scheme can be used for generating histograms for the unknown parameters from which, e.g., point and interval estimates can be derived. We also demonstrated the applicability of the proposed inference scheme to audio restoration. For a simple segment from a trumpet signal and a more complex segment from a speech signal, we recovered a 25 ms packet by use of the two neighbouring packets. Informal listening tests revealed that the restoration procedure restored the audio signal segments with a slight perceptual loss.

A Probability Distributions

In the following list, τ is a scalar positive random variable and $\boldsymbol{x}$ is an N-dimensional random vector.

Exponential Distribution

The exponential distribution with rate parameter λ has the probability distribution

$$p(\tau|\lambda) = \lambda \exp\{-\lambda\tau\}$$

and is denoted by $\mathrm{Exp}(\tau; \lambda)$.

Inverse Gamma Distribution

The inverse gamma distribution with shape parameter α and scale parameter β has the probability distribution

$$p(\tau|\alpha, \beta) = [\beta^\alpha/\Gamma(\alpha)]\tau^{-(\alpha+1)} \exp\{-\beta/\tau\}$$

and is denoted by $\text{Inv-}\mathcal{G}(\tau; \alpha, \beta)$.

Multivariate Normal Distribution

The multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ has the probability distribution

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left[(2\pi)^N |\boldsymbol{\Sigma}|\right]^{-1/2} \exp\{\tfrac{-1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}$$

and is denoted by $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Multivariate Normal-Scaled Inverse Gamma Distribution

The multivariate normal-scaled inverse gamma distribution with the location vector $\boldsymbol{\mu}$, covariance scale matrix $\boldsymbol{C}$, shape parameter α and scale parameter β has the probability distribution

$$p(\boldsymbol{x}, \tau|\boldsymbol{\mu}, \boldsymbol{C}, \alpha, \beta) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \tau\boldsymbol{C})\text{Inv-}\mathcal{G}(\tau; \alpha, \beta)$$

and is denoted by $\mathcal{NIG}(\boldsymbol{x}, \tau; \boldsymbol{\mu}, \boldsymbol{C}, \alpha, \beta)$.

Multivariate Student's t-Distribution

The multivariate student's t-distribution with the mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ and ν degrees of freedom has the probability distribution

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(N/2 + \nu/2)}{\Gamma(\nu/2)\sqrt{(\pi\nu)^N|\boldsymbol{\Sigma}|}} \left[1 + \frac{\Delta^2}{\nu}\right]^{-\frac{N+\nu}{2}}$$
$$\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$$

and is denoted by $\mathcal{T}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$.

Uniform Distribution

For $N = 1$, the uniform distribution with lower and upper boundary parameters a and b has the probability distribution

$$p(x|a, b) = \begin{cases} (b-a)^{-1} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

and is denoted by $\mathcal{U}(x; a, b)$.

References

[1] D. Goodman, G. Lockhart, O. Wasem, and W.-C. Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 6, pp. 1440–1448, Dec. 1986.

[2] Y. J. Liang, N. Farber, and B. Girod, "Adaptive playout scheduling and loss concealment for voice communication over IP networks," *IEEE Trans. Multimedia*, vol. 5, no. 4, pp. 532–543, Dec. 2003.

[3] Y. Wang, J. Li, and P. Stoica, *Spectral Analysis of Signals: The Missing Data Case*. Morgan and Claypool, Jun. 2005.

[4] H. Ofir, D. Malah, and I. Cohen, "Audio packet loss concealment in a combined MDCT-MDST domain," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 1032–1035, Dec. 2007.

[5] J. Rajan, P. Rayner, and S. Godsill, "Bayesian approach to parameter estimation and interpolation of time-varying autoregressive processes using the Gibbs sampler," *IEE Proc. Vis. Image Signal Process.*, vol. 144, no. 4, pp. 249–256, Aug. 1997.

[6] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration.* Springer-Verlag, London, 1998.

[7] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1609–1623, Sep. 2006.

[8] J. Lindblom, "A sinusoidal voice over packet coder tailored for the frame-erasure channel," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 787–798, Sep. 2005.

[9] C. A. Rødbro, M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Compressed domain packet loss concealment of sinusoidally coded speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Apr. 2003, pp. 104–107.

[10] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1625–1634, Jul. 2007.

[11] M. M. Goodwin, *Adaptive Signal Models: Theory, Algorithms and Audio Applications.* Springer, Oct. 1998.

[12] R. Kumaresan and D. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 19, no. 1, pp. 134–139, Jan. 1983.

[13] P. Stoica, R. L. Moses, B. Friedlander, and T. Söderström, "Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 3, pp. 378–392, Mar. 1989.

[14] J. Li and P. Stoica, "Efficient mixed-spectrum estimation with applications to target feature extraction," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, vol. 1, Oct. 1995, pp. 428–432.

[15] J. Cadzow, "Spectral estimation: An overdetermined rational model equation approach," *Proc. IEEE*, vol. 70, no. 9, pp. 907–939, Sep. 1982.

[16] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[17] A. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 8, Apr. 1983, pp. 336–339.

[18] A. Paulraj, R. Roy, and T. Kailath, "Estimation of signal parameters via rotational invariance techniques- ESPRIT," *Rec. Asilomar Conf. Signals, Systems, and Computers*, pp. 83–89, Nov. 1985.

[19] M. Viberg and B. Ottersten, "Sensor array processing based on subspace fitting," *IEEE Trans. Signal Process.*, vol. 39, no. 5, pp. 1110–1121, May 1991.

[20] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Prentice Hall, May 2005.

[21] J. M. Bernardo and A. Smith, *Bayesian Theory*, 1st ed. John Wiley and Sons Ltd, 1994.

[22] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Jun. 2002.

[23] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer-Verlag New York, Inc., Jul. 2004.

[24] E. T. Jaynes, "Bayesian spectrum and chirp analysis," in *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems*, C. R. Smith and G. J. Erickson, Eds. D. Reidel, Dordrecht-Holland, 1987, pp. 1–37.

[25] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*. Springer-Verlag, Berlin Heidelberg, 1988.

[26] L. Dou and R. J. W. Hodgson, "Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation I," *Inverse Problems*, vol. 11, no. 5, pp. 1069–1085, 1995.

[27] ——, "Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation II," *Inverse Problems*, vol. 12, no. 2, pp. 121–137, 1996.

[28] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2667–2676, 1999.

[29] P. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, pp. 711–732, 1995.

[30] D. V. Rubtsov and J. L. Griffin, "Time-domain Bayesian detection and estimation of noisy damped sinusoidal signals applied to NMR spectroscopy," *J. Magnetic Resonance*, vol. 188, no. 2, pp. 367–379, Aug. 2007.

[31] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of polyphonic western tonal music," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2498–2517, Apr. 2006.

[32] M. G. Christensen, S. V. Andersen, and S. H. Jensen, "Amplitude modulated sinusoidal models for audio modeling and coding," in *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 2773. Springer-Verlag, Oct. 2003, pp. 1334–1342.

[33] M. G. Christensen and S. van de Par, "Efficient parametric coding of transients," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1340–1351, Jul. 2006.

[34] S. J. Godsill and M. Davy, "Bayesian harmonic models for musical pitch estimation and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, 2002, pp. 1769–1772.

[35] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1989.

[36] A. C. Harvey and A. Jaeger, "Detrending, stylized facts and the business cycle," *J. Appl. Econometrics*, vol. 8, no. 3, pp. 231–47, Jul. 1993.

[37] A. Harvey, T. Trimbur, and H. v. Dijk, "Cyclical components in economic time series," Erasmus University Rotterdam, Econometric Institute, Econometric Institute Report EI 2002-20, Nov. 2002.

[38] R. Kleijn and H. K. van Dijk, "Bayes model averaging of cyclical decompositions in economic time series," *Journal of Applied Econometrics*, vol. 21, no. 2, pp. 191–212, Mar. 2006.

[39] A. T. Cemgil and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *Proc. European Signal Processing Conf.*, 2005.

[40] ——, "Efficient variational inference for the dynamic harmonic model," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2005, pp. 271–274.

[41] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[42] C. Dubois and M. Davy, "Joint detection and tracking of time-varying harmonic components: A flexible Bayesian approach." *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1283–1295, 2007.

[43] E. Vincent and M. D. Plumbley, "Efficient Bayesian inference for harmonic models via adaptive posterior factorization," *Neurocomput.*, vol. 72, no. 1-3, pp. 79–87, 2008.

[44] J. O Ruanaidh and W. Fitzgerald, "Interpolation of missing samples for audio restoration," *Electronics Letters*, vol. 30, no. 8, pp. 622–623, Apr. 1994.

[45] ——, *Numerical Bayesian Methods Applied to Signal Processing*, 1st ed. Springer-Verlag, New York, Feb. 1996.

[46] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Jul. 2003.

[47] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, Aug. 2006.

[48] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, Mar. 1953.

[49] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.

[50] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[51] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, November 1984.

[52] M. Stephens, "Dealing with label switching in mixture models," *J. Royal Stat. Soc., Series B*, vol. 62, no. 4, pp. 795–809, 2000.

[53] J. Durbin and S. J. Koopman, "A simple and efficient simulation smoother for state space time series analysis," *Biometrika*, vol. 89, no. 3, pp. 603–615, 2002.

[54] C. K. Carter and R. Kohn, "On Gibbs sampling for state space models," *Biometrika*, vol. 81, no. 3, pp. 541–553, Sep. 1994.

[55] P. De Jong and N. Shephard, "The simulation smoother for time series models," *Biometrika*, vol. 82, no. 2, pp. 339–350, Jun. 1995.

[56] J. Durbin and S. Koopman, *Time series analysis by state space methods*. Oxford University Press, 2001.

[57] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[58] L. E. Eberly and G. Casella, "Estimating Bayesian credible intervals," *J. Stat. Planning and Inference*, vol. 112, no. 1-2, pp. 115–132, Mar. 2003.

Paper D

Bayesian Interpolation in a Dynamic Sinusoidal Model with Application to Packet-Loss Concealment

Jesper Kjær Nielsen, Mads Græsbøll Christensen, Ali Taylan Cemgil, Simon John Godsill, and Søren Holdt Jensen

The paper has been published in the
Proceedings of the European Signal Processing Conference Aug. 2010.

The layout has been revised.

Abstract

In this paper, we consider Bayesian interpolation and parameter estimation in a dynamic sinusoidal model. This model is more flexible than the static sinusoidal model since it enables the amplitudes and phases of the sinusoids to be time-varying. For the dynamic sinusoidal model, we derive a Bayesian inference scheme for the missing observations, hidden states and model parameters of the dynamic model. The inference scheme is based on a Markov chain Monte Carlo method known as Gibbs sampler. We illustrate the performance of the inference scheme to the application of packet-loss concealment of lost audio and speech packets.

1 Introduction

Interpolation of missing, corrupted and future samples in signal waveforms is an important task in several applications. For example, speech and audio signals are often transmitted over packet-based networks in which packets may be lost, delayed or corrupted. If the contents of neighbouring packets are correlated, the erroneous packets can be approximately reconstructed by using suitable interpolation techniques. The simplest interpolation techniques employ signal repetition [1] and signal stretching [2], whereas more advanced interpolation techniques are based on filter bank methods such as GAPES and MAPES [3], and signal modelling such as autoregressive models [4, 5], hidden Markov models [6], and sinusoidal models [7]. An integral part of the techniques based on signal modelling is the estimation of the signal parameters. Given estimates of these parameters, the interpolation task is simply a question of simulating data from the model. In this paper, we develop an interpolation and parameter estimation scheme by assuming a dynamic sinusoidal model for an observed signal segment. This model can be written as a linear Gaussian time-invariant state space model given by

$$\begin{aligned} y_n &= \boldsymbol{b}^T \boldsymbol{s}_n + w_n \qquad \text{(observation equation)} \\ \boldsymbol{s}_{n+1} &= \boldsymbol{A}\boldsymbol{s}_n + \boldsymbol{v}_n \qquad \text{(state equation)} \end{aligned} \tag{D.1}$$

where $n = 1, \ldots, N$ label the uniform sampled data in time, and

$$\boldsymbol{b} = \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 \end{bmatrix}^T \tag{D.2}$$

$$\boldsymbol{A} = \mathrm{diag}(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_l, \cdots \boldsymbol{A}_L) \tag{D.3}$$

$$\boldsymbol{A}_l = \exp(-\gamma_l) \begin{bmatrix} \cos\omega_l & \sin\omega_l \\ -\sin\omega_l & \cos\omega_l \end{bmatrix}, \tag{D.4}$$

with $\omega_l \in [0, \pi]$ and $\gamma_l > 0$ denoting the (angular) frequency, and the log-damping coefficient of the l'th sinusoid, respectively. Further, $\boldsymbol{s}_n$ is the state vector, and $\boldsymbol{v}_n$ and w_n are white Gaussian state and observation noise sequences with covariance matrix

$\boldsymbol{Q}$ and variance σ_w^2, respectively. We also assume a Gaussian prior for the initial state vector $\boldsymbol{s}_1$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{P}$. For a non-zero state covariance matrix, the dynamic sinusoidal model in (D.1) is able to model non-stationary tonal signals such as a wide range of speech and audio signal segments. We are here concerned with the problem of performing interpolation and parameter estimation in the model in (D.1) from a Bayesian perspective which offer some conceptual advantages to classical statistics (see, e.g., [8]). For example, the Bayesian approach offers a standardised way of dealing with nuisance parameters and signal interpolation [4]. The downside of using the Bayesian methods is that they struggle with practical problems such as evaluation of high-dimensional and intractable integrals. Although various developments in Markov chain Monte Carlo (MCMC) methods (see, e.g., [9]) in recent years have overcome these problems to a great extend, the methods still remain very computational intensive.

Within the field of econometrics, the dynamic sinusoidal model in (D.1) is well-known and referred to as the stochastic cyclical model [10]. Two slightly different stochastic cyclical models were given a fully Bayesian treatment using MCMC inference techniques in [11] and [12]. Neither of these, however, considered the case where some observations are missing. In the audio and speech processing field, the dynamic sinusoidal model has also been considered by Cemgil et al. in [13–15]. However, they considered the frequency parameter as a discrete random variable and based their inference on approximate variational Bayesian methods.

In this paper, we extend the above work by developing an inference scheme for the dynamic sinusoidal model based on MCMC inference techniques. We consider the frequency parameter as a continuous random variable and allow some of the observations to be missing. To achieve this, we develop a Gibbs sampling scheme. The output of this sampler can be used for forming histograms of the unknown parameters of interest. These histograms have the desirable property that they converge to the probability distribution of these unknown parameters when the number of generated samples is increased, and they therefore enable us to extract statistical features for the model parameters as well as for performing the interpolation of the missing observations. It should be noted that although this inference scheme can be used for estimating parameters of signals with no missing observations, the primary focus of this paper is on the application of reconstructing missing observations from signal segments which are assumed to have been generated by a dynamic sinusoidal model.

The paper is organised as follows. In Sec. 2, we formalise the problem by setting up the Bayesian framework. This enables us in Sec. 3 to develop the interpolation and inference scheme. In Sec. 4, we illustrate the performance of the interpolating scheme by use of simulations, and Sec. 5 concludes this paper.

2 Problem Formulation

In the Bayesian approach, all variables of the model in (D.1) are random variables, and we partition them as

$$\begin{aligned} &\text{Observations:} && \boldsymbol{y} = \left[y_1, y_2, \cdots, y_N\right]^T \\ &\text{Latent variables:} && \boldsymbol{S} = \left[\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N\right] \\ &\text{Model parameters:} && \boldsymbol{\theta} = \{\boldsymbol{\omega}, \boldsymbol{\gamma}, \boldsymbol{q}, \sigma_w^2\} \end{aligned}$$

where $\boldsymbol{\omega}$, $\boldsymbol{\gamma}$ and $\boldsymbol{q}$ are L-dimensional vectors consisting of the L frequencies, the L log-damping parameters and the L state noise variances, respectively. The nth state vector $\boldsymbol{s}_n = \left[\boldsymbol{s}_{n,1}^T, \cdots, \boldsymbol{s}_{n,L}^T\right]^T$ consists of L two-dimensional state vectors pertaining to the L sinusoids. Conditioned on the previous state vector, each of these L two-dimensional state vectors has isotropic covariance matrix $q_l \boldsymbol{I}_2$, where $\boldsymbol{I}_2$ is the 2×2 identity matrix, so that $\boldsymbol{Q} = \text{diag}(\boldsymbol{q}) \otimes \boldsymbol{I}_2$ where $\otimes$ is the Kronecker product. We also assume that R of the elements in $\boldsymbol{y}$ are missing or corrupted, and that we know their indices $\mathcal{I} \subset \{1, \ldots, N\}$. Using this set of indices, we define the vectors $\boldsymbol{y}_\text{m} \triangleq \boldsymbol{y}_\mathcal{I}$ and $\boldsymbol{y}_\text{o} \triangleq \boldsymbol{y}_{\setminus\mathcal{I}}$ containing the R missing or corrupted observations and the $N - R$ valid observations, respectively. The notation $(\cdot)_{\setminus *}$ denotes 'without element $*$'.

The primary objective of this paper is to recover $\boldsymbol{y}_\text{m}$ from $\boldsymbol{y}_\text{o}$. This can be achieved in various ways, e.g., by using MAP/MMSE estimate w.r.t. the posterior distribution $p(\boldsymbol{y}_\text{m}|\boldsymbol{y}_\text{o})$ or by drawing a sample from $p(\boldsymbol{y}_\text{m}|\boldsymbol{y}_\text{o})$. The MAP-based interpolation produces the most probable interpolants. For audio and speech signals, however, MAP/MMSE-based interpolation tends to produce over-smoothed interpolants in the sense that they do not agree with the stochastic part of the valid observations [16]. A more typical interpolant can be obtained by drawing a single sample from $p(\boldsymbol{y}_\text{m}|\boldsymbol{y}_\text{o})$ [4]. The posterior distribution for the missing samples given the valid samples is given by

$$p(\boldsymbol{y}_\text{m}|\boldsymbol{y}_\text{o}) = \int p(\boldsymbol{y}_\text{m}|\boldsymbol{S}_\mathcal{I}, \sigma_w^2) p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_\text{o}) d\boldsymbol{S} d\boldsymbol{\theta} \; . \tag{D.5}$$

We are not able to draw a sample directly from $p(\boldsymbol{y}_\text{m}|\boldsymbol{y}_\text{o})$ since we are not able to integrate the nuisance parameters $\boldsymbol{S}$ and $\boldsymbol{\theta}$ out analytically. However, we can obtain a sample from $p(\boldsymbol{y}_\text{m}|\boldsymbol{y}_\text{o})$ by taking a single sample from the joint posterior distribution $p(\boldsymbol{y}_\text{m}, \boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_\text{o})$ and simply ignore the generated values for $\boldsymbol{S}$ and $\boldsymbol{\theta}$. From the observation equation of (D.1), we know the distribution of $p(\boldsymbol{y}_\text{m}|\boldsymbol{S}_\mathcal{I}, \sigma_w^2)$, so the only problem left is computing $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_\text{o})$. This distribution is by Bayes' theorem given by

$$p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_\text{o}) = \frac{p(\boldsymbol{y}_\text{o}, \boldsymbol{S}_{\setminus 1}|\boldsymbol{s}_1, \boldsymbol{\theta}) p(\boldsymbol{s}_1, \boldsymbol{\theta})}{p(\boldsymbol{y}_\text{o})} \tag{D.6}$$

where $p(\boldsymbol{y}_\text{o}, \boldsymbol{S}_{\setminus 1}|\boldsymbol{s}_1, \boldsymbol{\theta})$, $p(\boldsymbol{s}_1, \boldsymbol{\theta})$ and $p(\boldsymbol{y}_\text{o})$ are referred to as the likelihood, the prior and the model evidence, respectively. Under the above assumption, the likelihood can be

factored as

$$p(\boldsymbol{y}_{\mathrm{o}}, \boldsymbol{S}_{\setminus 1}|\boldsymbol{s}_1, \boldsymbol{\theta}) = p(\boldsymbol{y}_{\mathrm{o}}|\boldsymbol{S}_{\setminus \mathcal{I}}, \sigma_w^2) \times \prod_{n=1}^{N-1}\prod_{l=1}^{L} p(\boldsymbol{s}_{n+1,l}|\boldsymbol{s}_{n,l}, q_l, \omega_l, \gamma_l) \tag{D.7}$$

which from (D.1) is seen to be a product of normal distributions. For the prior distribution, we assume the factorisation

$$\begin{aligned} p(\boldsymbol{s}_1, \boldsymbol{\theta}) &= p(\boldsymbol{s}_1)p(\boldsymbol{\omega})p(\boldsymbol{\gamma})p(\boldsymbol{q})p(\sigma_w^2) \\ &= p(\boldsymbol{s}_1)\left[\prod_{l=1}^{L} p(\omega_l)p(\gamma_l)p(q_l)\right] p(\sigma_w^2) \end{aligned} \tag{D.8}$$

where $p(\boldsymbol{s}_1)$ has a normal distribution $\mathcal{N}(\boldsymbol{s}_1; \boldsymbol{\mu}, \boldsymbol{P})$, $p(\omega_l)$ has a uniform distribution $\mathcal{U}(\omega_l; 0, \pi)$, $p(\gamma_l)$ has an exponential distribution $\mathrm{Exp}(\gamma_l; \lambda_l)$, and $p(\sigma_w^2)$ and $p(q_l)$ have inverse gamma distributions Inv-$\mathcal{G}(\sigma_w^2; \alpha_w, \beta_w)$ and Inv-$\mathcal{G}(q_l; \alpha_{v,l}, \beta_{v,l})$. The model evidence $p(\boldsymbol{y}_{\mathrm{o}})$ is independent of $\boldsymbol{S}$ and $\boldsymbol{\theta}$ and is therefore a mere scale factor which can be ignored in the inference stage.

3 Inference Scheme

In the Bayesian framework, all statistical inference is based on the posterior distribution over the unknown variables or a marginal posterior distribution over some of these. As derived in the previous section, we have to generate samples from $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}})$ in order to be able to do this. Unfortunately, this distribution has a very complicated form, and we are therefore not able to sample directly from it. We therefore have to resort to other sampling techniques in order to enable statistical inference based on this distribution. One of the simplest and most popular numerical sampling techniques is the Gibbs sampler [17] which is an MCMC-based algorithm and suitable for this task. The Gibbs sampler draws samples from a multivariate distribution, say $p(\boldsymbol{x}) = p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K)$, by breaking it into a number of conditional distributions $p(\boldsymbol{x}_k|\boldsymbol{x}_{\setminus k})$ of smaller dimensionality from which samples are obtained in an alternating pattern. Specifically, for the τ's iteration, we sample for $k = 1, \ldots, K$ from

$$\boldsymbol{x}_k^{[\tau+1]} \sim p(\boldsymbol{x}_k|\boldsymbol{x}_1^{[\tau+1]}, \ldots, \boldsymbol{x}_{k-1}^{[\tau+1]}, \boldsymbol{x}_{k+1}^{[\tau]}, \ldots, \boldsymbol{x}_K^{[\tau]}) \ . \tag{D.9}$$

After an initial burn-in time during which the sampling scheme converges, the samples obtained from sampling these lower dimensional conditional distributions can be regarded as samples from the joint posterior distribution. In this paper, the posterior

distribution $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}})$ is broken into the two conditional distributions given by

$$\begin{aligned} &\text{States:} && p(\boldsymbol{S}|\boldsymbol{\theta}, \boldsymbol{y}_{\mathrm{o}}) && \text{(D.10)} \\ &\text{Model parameters:} && p(\boldsymbol{\theta}|\boldsymbol{S}, \boldsymbol{y}_{\mathrm{o}}) && \text{(D.11)} \end{aligned}$$

The selected grouping of variables in (D.10) and (D.11) leads to a set of conditional distributions which are fairly easy to sample from. In the next sections, we derive the particular form of these conditional distributions.

3.1 States

The conditional state distribution in (D.10) can be shown to be a multivariate Gaussian distribution. However, the dimension of this distribution is $2LN \times 1$ which would render direct sampling from it infeasible for most applications. Instead, we use the simulation smoother [18], which is an efficient sampling scheme using standard Kalman smoothing, for drawing samples from (D.10). Since some of the observations are missing, we have to modify the simulation smoother slightly. This is easily done by skipping the update step of the build-in Kalman filter for these observations.

3.2 Model Parameters

Since the model parameter of the observation equation, σ_w^2, and the L sets of model parameters of the state equation, $(\omega_l, \gamma_l, q_l)$, are mutually independent conditioned on the states $\boldsymbol{S}$, we can factor (D.11) as

$$p(\boldsymbol{\theta}|\boldsymbol{S}, \boldsymbol{y}_{\mathrm{o}}) = \left[\prod_{l=1}^{L} p(\omega_l, \gamma_l, q_l|\boldsymbol{S})\right] p(\sigma_w^2|\boldsymbol{S}, \boldsymbol{y}_{\mathrm{o}}) \ . \tag{D.12}$$

Thus, sampling from the conditional distribution in (D.11) can be done by sampling the $L + 1$ conditional distributions on the right side of (D.12) independently.

Frequency, Log-damping and State Noise Variance

To our knowledge, it is not possible to sample directly from the conditional distribution $p(\omega_l, \gamma_l, q_l|\boldsymbol{S})$. A Gibbs sampling scheme is also not straight-forward since it suffers from poor mixing and since the l'th log-damping coefficient conditioned on the l'th frequency parameter and state noise variance has a non-standard distribution. In order to improve mixing of the parameters and lower the overall computational complexity, we therefore propose sampling from $p(\omega_l, \gamma_l, q_l|\boldsymbol{S})$ by use of a Metropolis-Hastings (MH) sampler [19]. In the MH sampler, samples generated from the desired posterior distribution, say $p(\boldsymbol{x})$, which we know up to some normalising constant Z with $p(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x})/Z$, are generated by use of a user-defined proposal distribution $q(\boldsymbol{x}|\boldsymbol{x}^{[\tau]})$, where $\boldsymbol{x}^{[\tau]}$ is the τth generated

sample. In general, $p(\boldsymbol{x}) \neq q(\boldsymbol{x}|\boldsymbol{x}^{[\tau]})$ so a proposed sample $\boldsymbol{x}' \sim q(\boldsymbol{x}|\boldsymbol{x}^{[\tau]})$ is only accepted as a sample from $p(\boldsymbol{x})$ with probability

$$\alpha(\boldsymbol{x}^{[\tau]}, \boldsymbol{x}') = \min\left[1, \frac{\tilde{p}(\boldsymbol{x}')q(\boldsymbol{x}^{[\tau]}|\boldsymbol{x}')}{\tilde{p}(\boldsymbol{x}^{[\tau]})q(\boldsymbol{x}'|\boldsymbol{x}^{[\tau]})}\right] . \tag{D.13}$$

Otherwise, the previous accepted sample is retained, i.e., $\boldsymbol{x}^{[\tau+1]} = \boldsymbol{x}^{[\tau]}$.

For $p(\omega_l, \gamma_l, q_l|\boldsymbol{S})$, the proposal samples $(\omega_l', \gamma_l', q_l')$ are generated in two simple steps: First, we generate a sample for the mean and variance of a bivariate normal-scaled inverse gamma distribution with isotropic covariance matrix. This is done by sampling from

$$q_l' \sim \text{Inv-}\mathcal{G}(\alpha_{q_l}, \beta_{q_l}) \tag{D.14}$$
$$\tau_l' \sim \text{Inv-}\mathcal{G}(\alpha_{q_l}, 1/2) \tag{D.15}$$
$$\boldsymbol{a}_l' = \begin{bmatrix} a_{1,l}' & a_{2,l}' \end{bmatrix}^T \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{a},l}, 2\beta_{q_l}\tau_l'\sigma_{\boldsymbol{a},l}^2\boldsymbol{I}_2) \tag{D.16}$$

where we have defined

$$\boldsymbol{\varphi}_l \triangleq \begin{bmatrix} \boldsymbol{s}_{2,l}^T & \boldsymbol{s}_{3,l}^T & \cdots & \boldsymbol{s}_{N,l}^T \end{bmatrix}^T \tag{D.17}$$
$$\boldsymbol{\phi}_l \triangleq \begin{bmatrix} \boldsymbol{s}_{1,l}^T & \boldsymbol{s}_{2,l}^T & \cdots & \boldsymbol{s}_{N-1,l}^T \end{bmatrix}^T \tag{D.18}$$
$$\tilde{\boldsymbol{\phi}}_l \triangleq \begin{bmatrix} (\boldsymbol{s}_{1,l}^{\perp})^T & (\boldsymbol{s}_{2,l}^{\perp})^T & \cdots & (\boldsymbol{s}_{N-1,l}^{\perp})^T \end{bmatrix}^T \tag{D.19}$$
$$\boldsymbol{\Phi}_l \triangleq \begin{bmatrix} \boldsymbol{\phi}_l & \tilde{\boldsymbol{\phi}}_l \end{bmatrix} \tag{D.20}$$
$$\sigma_{\boldsymbol{a},l}^2 \triangleq (\boldsymbol{\phi}_l^T\boldsymbol{\phi}_l)^{-1} \tag{D.21}$$
$$\boldsymbol{\mu}_{\boldsymbol{a},l} \triangleq \sigma_{\boldsymbol{a},l}^2\boldsymbol{\Phi}_l^T\boldsymbol{\varphi}_l \tag{D.22}$$
$$\alpha_{q_l} \triangleq \alpha_{v,l} + N - 1 \tag{D.23}$$
$$\beta_{q_l} \triangleq \beta_{v,l} + (\boldsymbol{\varphi}_l^T\boldsymbol{\varphi}_l - \sigma_{\boldsymbol{a},l}^{-2}\boldsymbol{\mu}_{\boldsymbol{a},l}^T\boldsymbol{\mu}_{\boldsymbol{a},l})/2 , \tag{D.24}$$

and $\boldsymbol{s}_{n,l}^{\perp}$ is obtained by a 90° clockwise rotation of $\boldsymbol{s}_{n,l}$. Second, we transform $\boldsymbol{a}_l'$ into (ω_l', γ_l') by the relations

$$\omega_l' = \arctan(a_{2,l}'/a_{1,l}') \tag{D.25}$$
$$\gamma_l' = -\ln\left(\boldsymbol{a}_l'^T\boldsymbol{a}_l'\right)/2 . \tag{D.26}$$

Then, if $a_{2,l}' \geq 0$, the proposal samples $(\omega_l', \gamma_l', q_l')$ are accepted as samples from $p(\omega_l, \gamma_l, q_l|\boldsymbol{S})$ with probability

$$\alpha\big((\omega_l^{[\tau]}, \gamma_l^{[\tau]}, q_l^{[\tau]}), (\omega_l', \gamma_l', q_l')\big) = \min\left[1, \exp\left\{(\lambda_l - 2)(\gamma_l^{[\tau]} - \gamma_l')\right\}\right] . \tag{D.27}$$

Otherwise, the previous values $(\omega_l^{[\tau]}, \gamma_l^{[\tau]}, q_l^{[\tau]})$ are retained. Notice that if the rate parameter λ_l of the prior for γ_l is equal to two, $\alpha = 1$ for any γ_l'. The details of the derivation of this sampling scheme can be found in [20].

Observation Noise Variance

By Bayes' theorem, we can write $p(\sigma_w^2|\boldsymbol{S}, \boldsymbol{y}_\mathrm{o})$ as

$$p(\sigma_w^2|\boldsymbol{S}, \boldsymbol{y}_\mathrm{o}) \propto p(\boldsymbol{y}_\mathrm{o}|\boldsymbol{S}_{\setminus\mathcal{I}}, \sigma_w^2)p(\sigma_w^2) \tag{D.28}$$

where $p(\boldsymbol{y}_\mathrm{o}|\boldsymbol{S}_{\setminus\mathcal{I}}, \sigma_w^2)$ is the likelihood of the observation equation in (D.1) and $p(\sigma_w^2)$ is the prior distribution for σ_w^2. Since $p(\boldsymbol{y}_\mathrm{o}|\boldsymbol{S}_{\setminus\mathcal{I}}, \sigma_w^2) = \mathcal{N}(\boldsymbol{S}_{\setminus\mathcal{I}}^T\boldsymbol{b}, \sigma_w^2\boldsymbol{I}_{N-R})$ and $p(\sigma_w^2) = \text{Inv-}\mathcal{G}(\alpha_w, \beta_w)$, the posterior distribution $p(\sigma_w^2|\boldsymbol{S}_{\setminus\mathcal{I}}, \boldsymbol{y}_\mathrm{o})$ is an inverse gamma distribution, $\text{Inv-}\mathcal{G}(\sigma_w^2; \alpha_{\sigma_w^2}, \beta_{\sigma_w^2})$, with parameters

$$\alpha_{\sigma_w^2} = \alpha_w + N/2 \tag{D.29}$$

$$\beta_{\sigma_w^2} = \beta_w + \frac{1}{2}(\boldsymbol{y}_\mathrm{o} - \boldsymbol{S}_{\setminus\mathcal{I}}^T\boldsymbol{b})^T(\boldsymbol{y}_\mathrm{o} - \boldsymbol{S}_{\setminus\mathcal{I}}^T\boldsymbol{b}) \ . \tag{D.30}$$

3.3 Summary of Inference Scheme

Table D.1 summarises our proposed Gibbs sampler for generating samples from $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_\mathrm{o})$. The computational complexity of the algorithm is fairly high primarily due to the generation of the states by the simulation smoother. In our implementation with $N = 600$ observations and $L = 6$ sinusoids, it takes approximately 40 ms for generating a state sample $\boldsymbol{S}^{[\tau]}$. This corresponds to nearly 97 % of the time consumption of one iteration of the Gibbs sampler. For the application of interpolation, we only need a single sample for the states and model parameters from the invariant distribution of the underlying Markov chain of the sampler. Once these have been generated, we may perform the interpolation by simulating from the observation equation of (D.1). Therefore, the computational complexity of the algorithm heavily depends on proper initialisation and the convergence speed of the chain.

4 Simulations

We consider the problem of reconstructing missing or corrupted packets on a packet-based network. First, we illustrate the reconstruction process and, second, we present the results of a small-scale listening test.

1. Select hyperparameters and initialise the Gibbs sampler.
2. Repeat for $k = 0, 1, 2, \ldots, K$
   (a) $\boldsymbol{S}^{[k+1]} \sim p(\boldsymbol{S}|\boldsymbol{\theta}^{[k]}, \boldsymbol{y}_{\mathrm{o}})$ (simulation smoother)
   (b) Repeat for $l = 1, 2, \ldots, L$
      i. $q_l' \sim \text{Inv-}\mathcal{G}(\alpha_{q_l}^{[\tau]}, \beta_{q_l}^{[\tau]})$
      ii. $\tau_l' \sim \text{Inv-}\mathcal{G}(\alpha_{q_l}^{[\tau]}, 1/2)$
      iii. $\boldsymbol{a}_l' \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{a},l}^{[\tau]}, 2\beta_{q_l}^{[\tau]}\tau_l'\sigma_{\boldsymbol{a},l}^{2\ [\tau]}\boldsymbol{I}_2)$
      iv. $\omega_l' = \arctan(a_{2,l}'/a_{1,l}')$
      v. $\gamma_l' = -\ln\left(\boldsymbol{a}_l'^T\boldsymbol{a}_l'\right)/2$
      vi. $u_l = \mathcal{U}(0,1)$
      vii. if $u_l \leq \alpha\big((\omega_l^{[\tau]}, \gamma_l^{[\tau]}, q_l^{[\tau]}), (\omega_l', \gamma_l', q_l')\big)$
         - $(\omega_l^{[k+1]}, \gamma_l^{[k+1]}, q_l^{[k+1]}) = (\omega_l', \gamma_l', q_l')$

         else
         - $(\omega_l^{[k+1]}, \gamma_l^{[k+1]}, q_l^{[k+1]}) = (\omega_l^{[\tau]}, \gamma_l^{[\tau]}, q_l^{[\tau]})$

   (c) $\sigma_w^{2\ [k+1]} \sim \text{Inv-}\mathcal{G}(\alpha_{\sigma_w^2}^{[\tau]}, \beta_{\sigma_w^2}^{[\tau]})$

Table D.1: Summary of proposed Gibbs sampler for generating samples from $p(\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{y}_{\mathrm{o}})$.

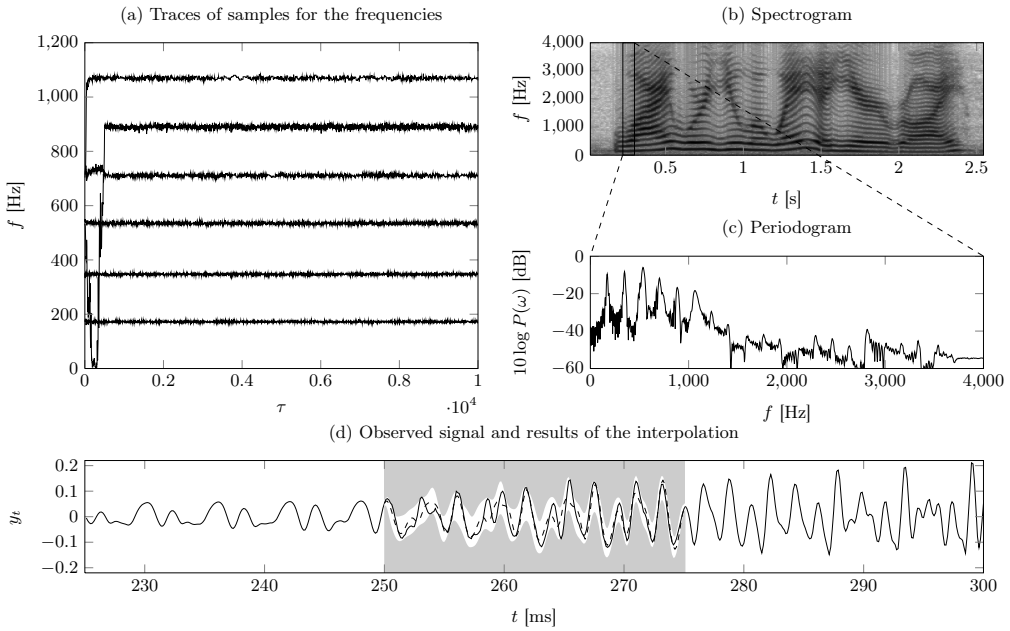

Fig. D.1: Plot (a) shows the six traces for the frequencies each consisting of 10,000 samples. Plot (b) shows the spectrogram for the complete speech signal whereas plot (c) shows the periodogram for the section indicated in plot (b). The time series corresponding to this section is shown in plot (d) with the middle section of 25 ms audio missing. The plot also shows the result of the interpolation in terms of the 95 % credible interval, a sample from the marginal posterior distribution $p(\boldsymbol{y}_{\mathrm{m}}|\boldsymbol{y}_{\mathrm{o}})$ and the true missing observations (dashed).

4.1 Speech Signal Reconstruction

We used a snapshot from a speech signal (see Fig. D.1.d) consisting of $N = 600$ samples corresponding to 75 ms of speech at a sampling frequency of 8000 kHz. The speech signal is generated by a female voice uttering, *"Why were you away a year, Roy?"* and its spectrogram in shown in Fig. D.1.b. The periodogram of the 75 ms speech signal segment is shown in Fig. D.1.c. Prior to running the Gibbs sampler, we removed the middle section thus emulating a lost audio packet of 25 ms. For the setup of the Gibbs sampler, we assumed $L = 6$ sinusoidal components, and we selected the hyperparameters such that the prior distributions were diffuse. The initial values for the frequency and the observation noise variance were computed by using a matching pursuit algorithm. The initial values for the log-damping coefficients and the state noise covariances were somewhat heuristically set to 0 and ${\sigma_w^2}^{[0]}/10$, respectively. Fig. D.1 shows the main results of the simulation. Fig. D.1.a shows the six traces of samples

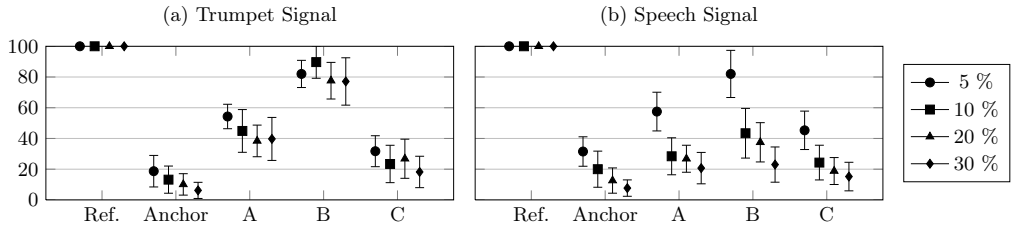

Fig. D.2: Mean and 95 % confidence intervals for the MUSHRA listening test. The reference signal was transmitted through four artificial channels with independent packet-loss probabilities of 5 %, 10 %, 20 % and 30 %, respectively. For the anchor signal, the missing packets were interpolated with zeros while the interpolation for A, B and C were based on $p(\boldsymbol{y}_\mathrm{m}|\boldsymbol{y}_\mathrm{o})$, $E\{\boldsymbol{y}_\mathrm{m}|\boldsymbol{y}_\mathrm{o}\}$ and $p(\boldsymbol{y}_\mathrm{m}|\boldsymbol{y}_\mathrm{o}, \boldsymbol{\theta}^{\mathrm{MAP}})$, respectively.

obtained for the frequency parameters. After a burn-in length of approximately 1000 samples the underlying Markov chain seems to have converged to the true posterior distribution for the frequencies. Inference for the frequency parameters can thus be based on histograms formed by the the last approximately 9000 samples. In a similar way, histograms for the remaining model parameters can be formed. Fig. D.1.d shows a typical sample obtained for the missing observations compared to the true signal. Notice, that unlike maximum likelihood- and EM-restoration techniques, the noise is also modelled when performing the interpolation in the Bayesian framework. Fig. D.1.d also shows an estimate of the 95 % credible interval for the missing observations.

4.2 Listening Test

We conducted a small-scale MUSHRA listening test [21, 22] to evaluate the performance of the interpolation scheme. In addition to the speech signal, we also used an excerpt from a trumpet signal. Both of these signals were partitioned in 25 ms packets and transmitted through four artificial channels where packets were lost independently with probabilities of 5 %, 10 %, 20 % and 30 %, respectively. On the receiver side, we applied our proposed interpolation scheme to the missing packets. For every gap of one or more consecutive missing packets, we used the valid packet before and after the gap as in Fig. D.1. We compared the interpolant (A) from $p(\boldsymbol{y}_\mathrm{m}|\boldsymbol{y}_\mathrm{o})$ against the MMSE interpolant (B) $E\{\boldsymbol{y}_\mathrm{m}|\boldsymbol{y}_\mathrm{o}\}$ and the interpolant (C) from $p(\boldsymbol{y}_\mathrm{m}|\boldsymbol{y}_\mathrm{o}, \boldsymbol{\theta}^{\mathrm{MAP}})$. The MMSE and MAP estimates were computed from the last 9000 generated samples from the Gibbs sampler. For the anchor signal, we used zeros for the interpolation. Fig. D.2 shows the results obtained by applying the statistical analysis suggested in [21] to the scores given by ten listeners. The listening test clearly revealed that reconstructing missing packets of the highly tonal and fairly stationary trumpet signal was much more successful than for the speech signal. The results also revealed that the interpolation based on $E\{\boldsymbol{y}_\mathrm{m}|\boldsymbol{y}_\mathrm{o}\}$ performed better than the other methods.

5 Conclusion

Based on a Gibbs sampler, we have presented a Bayesian inference scheme for the missing observations, the states and the model parameters of a dynamic sinusoidal model. This model is able to model some non-stationary signal segments which are often encountered in music or speech signal processing. In the simulations, we demonstrated that the algorithm can be used for interpolation of audio and speech signals. This is an integral part of many signal processing applications such as packet-loss concealment, pitch- and time-scale modification. Additionally, the inference scheme can also be used for making inference about the unknown model parameters of the dynamic sinusoidal model.

References

[1] D. Goodman, G. Lockhart, O. Wasem, and W.-C. Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 6, pp. 1440–1448, Dec. 1986.

[2] Y. J. Liang, N. Farber, and B. Girod, "Adaptive playout scheduling and loss concealment for voice communication over IP networks," *IEEE Trans. Multimedia*, vol. 5, no. 4, pp. 532–543, Dec. 2003.

[3] H. Ofir and D. Malah, "Packet loss concealment for audio streaming based on the GAPES and MAPES algorithms," in *IEEE Conv. Electrical and Electronics Engineers in Isreal*, Nov. 2006, pp. 280–284.

[4] S. J. Godsill and P. J. W. Rayner, *Digital Audio Restoration.* Springer-Verlag, London, 1998.

[5] M. Lagrange, S. Marchand, and J.-B. Rault, "Enhancing the tracking of partials for the sinusoidal modeling of polyphonic sounds," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1625–1634, Jul. 2007.

[6] C. A. Rødbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1609–1623, Sep. 2006.

[7] J. Lindblom and P. Hedelin, "Packet loss concealment based on sinusoidal modeling," in *IEEE Proc. Workshop on Speech Coding*, Oct. 2002, pp. 65–67.

[8] J. M. Bernardo and A. Smith, *Bayesian Theory*, 1st ed. John Wiley and Sons Ltd, 1994.

[9] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed. Springer-Verlag New York, Inc., Jul. 2004.

[10] A. C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1989.

[11] A. Harvey, T. Trimbur, and H. v. Dijk, "Cyclical components in economic time series," Erasmus University Rotterdam, Econometric Institute, Econometric Institute Report EI 2002-20, Nov. 2002.

[12] R. Kleijn and H. K. van Dijk, "Bayes model averaging of cyclical decompositions in economic time series," *Journal of Applied Econometrics*, vol. 21, no. 2, pp. 191–212, Mar. 2006.

[13] A. T. Cemgil and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals," in *Proc. European Signal Processing Conf.*, 2005.

[14] ——, "Efficient variational inference for the dynamic harmonic model," in *Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2005, pp. 271–274.

[15] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[16] J. O Ruanaidh and W. Fitzgerald, "Interpolation of missing samples for audio restoration," *Electronics Letters*, vol. 30, no. 8, pp. 622–623, Apr. 1994.

[17] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, November 1984.

[18] J. Durbin and S. J. Koopman, "A simple and efficient simulation smoother for state space time series analysis," *Biometrika*, vol. 89, no. 3, pp. 603–615, 2002.

[19] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.

[20] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, S. J. Godsill, and S. H. Jensen, "Bayesian interpolation and parameter estimation in a dynamic sinusoidal model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1986–1998, Sep. 2011.

[21] "Method for the subjective assessment of intermediate quality levels of coding systems," 2003, ITU BS.1534-1.

[22] E. Vincent, *MUSHRAM: A MATLAB interface for MUSHRA listening tests*, 2005. [Online]. Available: http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/

Paper E

An Amplitude Spectral Capon Estimator with a Variable Filter Length

Jesper Kjær Nielsen, Paris Smaragdis, Mads Græsbøll Christensen, and Søren Holdt Jensen

The paper has been accepted for publication in the *Proceedings of the European Signal Processing Conference* Aug. 2012.

The layout has been revised.

Abstract

The filter bank methods have been a popular non-parametric way of computing the complex amplitude spectrum. So far, the length of the filters in these filter banks has been set to some constant value independently of the data. In this paper, we take the first step towards considering the filter length as an unknown parameter. Specifically, we derive a very simple and approximate way of determining the optimal filter length in a data-adaptive way. Based on this analysis, we also derive a model averaged version of the forward and the forward-backward amplitude spectral Capon estimators. Through simulations, we show that these estimators significantly improve the estimation accuracy compared to the traditional Capon estimators.

1 Introduction

The estimation of the complex amplitude spectrum is an important problem in several applications such as audio coding and radar imaging (see [1] and the references therein). Several solutions have been proposed in the literature ranging from simple estimators based on various windowed Fourier transforms to more complex estimators based on a parametric model of the observed data. A popular example of such a parametric model is the sinusoidal model

$$x(n) = \sum_{i=1}^{l} \alpha_i \exp(j\omega_i n) + w(n) \ , \quad n = 1, 2, \cdots, N \tag{E.1}$$

where α_i and ω_i are the complex amplitude[1] and the frequency of the i'th complex sinusoid, respectively, and $w(n)$ is a stationary random process with a possibly non-flat power spectral density (psd). Unfortunately, the parametric methods are usually very sensitive to modelling errors such as the noise statistics [2] and may also suffer from a high computational complexity if the model has non-linear parameters such as the frequency. Therefore, non-parametric methods may yield much better estimation results or lower the computational complexity significantly. For the estimation of the complex amplitude spectrum, the filter bank methods such as Capon [3] and APES [4] are examples of such non-parametric estimators which are fast and very robust to modelling errors. Although the APES method was originally proposed as an approximate maximum likelihood method, it can also be interpreted as a matched filter bank method like the Capon method [5]. Under this interpretation, the observed data is passed through an m-tap FIR-filter which is designed to maximise the signal-to-noise ratio (SNR) of the filter output subject to the constraint that the filter has a gain of one at some known frequency ω. The complex amplitude α at this frequency is then estimated from the filter

[1] Note that α_i is not an amplitude in the usual sense since it is not a real, positive scaler. However, in the lack of better words, we refer to it as the complex amplitude.

output. By designing a filter for all desired frequency points, we get a filter bank whose outputs are used to estimate the complex amplitudes at all these frequency points. The statistical properties of the Capon and APES methods have been studied extensively in, e.g., [5–7]. These studies have shown that the APES amplitude estimates are unbiased for all filter lengths, but that the Capon estimator gives amplitude estimates which are biased towards zero. For long filter lengths, this bias increases significantly. On the other hand, the Capon method has in general a better resolution than APES and is therefore more practical for estimating frequencies [1]. In [4], the maximum filter length of $m = \lfloor N/2 \rfloor$ is recommended for the APES method since this maximises the resolution. For the Capon method, however, there is a trade-off between resolution and bias, and in [7] a filter length in the interval $N/8 < m < N/4$ is recommended. To the best of our knowledge, the filter length m is always selected to be the same for all frequency points and does not depend on the observed data. In this paper, we take a first step towards determining the optimal filter length for the Capon method in a data-adaptive way. Specifically, we give a simple and approximate solution which is based on Djuric's asymptotic MAP approach [8].

2 The Amplitude Spectral Capon Estimator

The filter bank methods are a way of bypassing some of the difficulties associated with the parametric methods. This is achieved by first rewriting (E.1) as

$$x(n) = \alpha \exp(j\omega n) + z(n) \ , \quad n = 1, 2, \cdots, N \tag{E.2}$$

where α is the complex amplitude at the known (angular) frequency ω. In practice, we rarely know the true frequency parameters $\{\omega_i\}_{i=1}^{l}$ or the number l of them in (E.1), and these quantities are typically hard to estimate. In the filter bank methods, this problem is bypassed by selecting a set Ω of R candidate frequencies for which we wish to estimate the complex amplitude α. Comparing (E.2) to (E.1), we see that $z(n)$ models all the l sinusoids and the coloured noise $w(n)$, except for the complex sinusoid with an unknown complex amplitude α at the known frequency ω. We then pass this signal through an m-tap FIR filter and obtain

$$y(n) = \alpha \exp(j\omega n)\boldsymbol{h}^H \boldsymbol{a} + \boldsymbol{h}^H \boldsymbol{z}(n) \ . \tag{E.3}$$

for $n = m, \cdots, N$ where we have defined

$$\boldsymbol{a} \triangleq \begin{bmatrix} 1 & \exp(-j\omega) & \cdots & \exp(-j\omega(m-1)) \end{bmatrix}^T \tag{E.4}$$

$$\boldsymbol{h} \triangleq \begin{bmatrix} h_0 & h_1 & \cdots & h_{m-1} \end{bmatrix}^H \tag{E.5}$$

$$\boldsymbol{z}(n) \triangleq \begin{bmatrix} z(n) & z(n-1) & \cdots & z(n-m+1) \end{bmatrix}^T \ . \tag{E.6}$$

The notation $(\cdot)^T$ and $(\cdot)^H$ denote transposition and complex transposition, respectively. Maximising the SNR of the filter output is equivalent to solving [5]

$$\arg\min_{\boldsymbol{h}\in\mathbb{C}^m} \boldsymbol{h}^H\boldsymbol{Q}\boldsymbol{h} \quad \text{subject to} \quad \boldsymbol{h}^H\boldsymbol{a} = 1 \tag{E.7}$$

whose solution is

$$\boldsymbol{h} = (\boldsymbol{a}^H\boldsymbol{Q}^{-1}\boldsymbol{a})^{-1}\boldsymbol{Q}^{-1}\boldsymbol{a} \ . \tag{E.8}$$

The covariance matrix $\boldsymbol{Q} \triangleq E\{\boldsymbol{z}(n)\boldsymbol{z}^H(n)\}$ is unknown and must be estimated in some way. For $\boldsymbol{C} \triangleq E\{\boldsymbol{x}(n)\boldsymbol{x}^H(n)\}$ where $\boldsymbol{x}(n)$ is defined analogously to $\boldsymbol{z}(n)$, we have that

$$\boldsymbol{C} = |\alpha|^2\boldsymbol{a}\boldsymbol{a}^H + \boldsymbol{Q} \ , \tag{E.9}$$

and a simple estimate of $\boldsymbol{Q}$ is therefore $\hat{\boldsymbol{Q}} = \hat{\boldsymbol{C}} - |\hat{\alpha}|^2\boldsymbol{a}\boldsymbol{a}^H$. Inserting this estimate in (E.7) yields

$$\arg\min_{\boldsymbol{h}\in\mathbb{C}^m} \boldsymbol{h}^H\hat{\boldsymbol{C}}\boldsymbol{h} \quad \text{subject to} \quad \boldsymbol{h}^H\boldsymbol{a} = 1 \tag{E.10}$$

so that the Capon filter is

$$\boldsymbol{h}^{\text{Capon}} = (\boldsymbol{a}^H\hat{\boldsymbol{C}}^{-1}\boldsymbol{a})^{-1}\hat{\boldsymbol{C}}^{-1}\boldsymbol{a} \ . \tag{E.11}$$

The covariance matrix $\boldsymbol{C}$ of the input vector $\boldsymbol{x}(n)$ is typically estimated in one of two different ways. If we define $K \triangleq N - m + 1$ and

$$\boldsymbol{X} \triangleq \begin{bmatrix} \boldsymbol{x}(m) & \cdots & \boldsymbol{x}(N) \end{bmatrix} \ , \tag{E.12}$$

the forward estimate is given by

$$\hat{\boldsymbol{C}}_{\text{f}} = K^{-1}\boldsymbol{X}\boldsymbol{X}^H \ , \tag{E.13}$$

and the forward-backward (FB) estimate is given by

$$\hat{\boldsymbol{C}}_{\text{fb}} = (\hat{\boldsymbol{C}}_{\text{f}} + \boldsymbol{J}_m\hat{\boldsymbol{C}}_{\text{f}}^T\boldsymbol{J}_m)/2 \tag{E.14}$$

where $\boldsymbol{J}_m$ is the $m \times m$ exchange matrix. Like the true covariance matrix, $\hat{\boldsymbol{C}}_{\text{fb}}$ is persymmetric and simulation results have shown that it reduces the bias of the complex amplitude estimate significantly compared to $\hat{\boldsymbol{C}}_{\text{f}}$ [6]. However, the resolution is slightly better for $\hat{\boldsymbol{C}}_{\text{f}}$ [9]. The APES filter can be derived by using a different estimate of $\boldsymbol{Q}$, which can be found in [5], and it also exists in a forward and a FB version [6].

Due to the constraint $\boldsymbol{h}^H\boldsymbol{a} = 1$, we can write the filter output in (E.3) in vector form as

$$\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{h}^* = \alpha\boldsymbol{b} + \boldsymbol{e} \tag{E.15}$$

where $(\cdot)^*$ denotes complex conjugation and

$$\boldsymbol{y} \triangleq \begin{bmatrix} y(m) & \cdots & y(N) \end{bmatrix}^T \tag{E.16}$$

$$e(n) \triangleq \boldsymbol{h}^H \boldsymbol{z}(n) \tag{E.17}$$

$$\boldsymbol{e} \triangleq \begin{bmatrix} e(m) & \cdots & e(N) \end{bmatrix}^T \tag{E.18}$$

$$\boldsymbol{b} \triangleq \begin{bmatrix} \exp(j\omega m) & \cdots & \exp(j\omega N) \end{bmatrix}^T . \tag{E.19}$$

The least squares estimate of the complex amplitude is then

$$\hat{\alpha}_m = K^{-1} \boldsymbol{h}^H \boldsymbol{X} \boldsymbol{b}^* \tag{E.20}$$

where we have used the subscript m to indicate the length of the filter. When the Capon filter is used in (E.20), we term the resulting estimator as the amplitude spectral Capon (ASC) estimator.

3 The Model Averaged ASC Estimator

To estimate the optimal filter length, we first briefly review the asymptotic MAP approach by Djuric [8]. In his framework, we wish to find the posterior distribution $p(m|\boldsymbol{x})$ on the filter length m given the K data points in the vector $\boldsymbol{x}$. This distribution is by Bayes' theorem given by

$$p(m|\boldsymbol{x}) \propto p(\boldsymbol{x}|m)p(m) \tag{E.21}$$

where $\propto$ denotes 'proportional to', and $p(\boldsymbol{x}|m)$ is referred to as the model likelihood or evidence which is given by

$$p(\boldsymbol{x}|m) = \int_{\Theta_m} p(\boldsymbol{x}|\boldsymbol{\theta}_m, m)p(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m \tag{E.22}$$

where $\boldsymbol{\theta}_m$ denotes the d_m model parameters with support Θ_m, $p(\boldsymbol{x}|\boldsymbol{\theta}_m, m)$ is the likelihood, and $p(\boldsymbol{\theta}_m|m)$ is the prior on the model parameters under model index m. Unfortunately, the integral in (E.22) can in general not be evaluated analytically, so Djuric suggests that the integral is approximately evaluated using the Laplace approximation. Provided that $\boldsymbol{\theta}$ is purely complex, the Laplace approximation gives

$$p(\boldsymbol{x}|m) \approx f(\hat{\boldsymbol{\theta}}_m)\pi^{d_m} |-\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m)|^{-1} \tag{E.23}$$

where $f(\boldsymbol{\theta}_m) \triangleq p(\boldsymbol{x}|\boldsymbol{\theta}_m, m)p(\boldsymbol{\theta}_m|m)$ is the integrand of (E.22), $\hat{\boldsymbol{\theta}}_m$ is the MAP estimate of $\boldsymbol{\theta}_m$, and

$$\boldsymbol{H}(\boldsymbol{\theta}_m) = \frac{\partial^2 \ln f(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m^* \partial \boldsymbol{\theta}_m^T} \tag{E.24}$$

is the Hessian matrix. Djuric also suggests that we simplify (E.23) by neglecting the terms of order $\mathcal{O}(1)$ and by evaluating the determinant of the observed information matrix $-\boldsymbol{H}(\hat{\boldsymbol{\theta}}_m)$ using asymptotic considerations. Specifically, for the linear Gaussian model $\boldsymbol{x} = \boldsymbol{F}_m\boldsymbol{\theta}_m + \boldsymbol{e}$ and for a uniform prior on the model index and the model parameters, Djuric approximates $p(m|\boldsymbol{x})$ on the model index by [8]

$$p(m|\boldsymbol{x}) \propto (\hat{\sigma}^2)^{-K}|\boldsymbol{F}_m^H\boldsymbol{F}_m|^{-1} \tag{E.25}$$

where $\hat{\sigma}^2$ is the maximum likelihood estimate of the noise variance. The asymptotic approximation of $|\boldsymbol{F}_m^H\boldsymbol{F}_m|$ depends on the particular structure of $\boldsymbol{F}_m$.

3.1 Derivation

To construct a simple way of selecting the optimal filter length, we first integrate the constraint $\boldsymbol{h}^H\boldsymbol{a} = 1$ into the filter vector. Specifically, we write $h_0 = 1 - \tilde{\boldsymbol{a}}^H\tilde{\boldsymbol{h}}$ with

$$\tilde{\boldsymbol{h}} \triangleq \begin{bmatrix} h_1 & h_2 & \cdots & h_{m-1} \end{bmatrix}^T \tag{E.26}$$

$$\tilde{\boldsymbol{a}} \triangleq \begin{bmatrix} \exp(j\omega) & \exp(j\omega 2) & \cdots & \exp(j\omega m) \end{bmatrix}^T \tag{E.27}$$

so that the constraint $\boldsymbol{h}^H\boldsymbol{a} = 1$ is satisfied for all $\tilde{\boldsymbol{h}}$. This leads to that

$$\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{h}^* = \boldsymbol{x}_1 - \boldsymbol{F}_m\tilde{\boldsymbol{h}} \tag{E.28}$$

where we have defined

$$\boldsymbol{x}_1 \triangleq \begin{bmatrix} x(m) & \cdots & x(N) \end{bmatrix}^T \tag{E.29}$$

$$\boldsymbol{F}_m \triangleq \boldsymbol{X}^T \begin{bmatrix} \tilde{\boldsymbol{a}}^H \\ -\boldsymbol{I}_{m-1} \end{bmatrix} . \tag{E.30}$$

where $\boldsymbol{I}_{m-1}$ is the $(m-1)$-dimensional identity matrix. We now assume that $\boldsymbol{x}_1$ is independent of $\boldsymbol{F}_m$ and that $\boldsymbol{y}$ is a white Gaussian noise vector, so that $\boldsymbol{x}_1 = \boldsymbol{F}_m\tilde{\boldsymbol{h}} + \boldsymbol{y}$ is our simple linear model. Although this is in direct contradiction with the model of the filter output in (E.15), it is a necessary assumption to derive the ASC estimator in our framework. The APES estimator would be obtained instead if $\boldsymbol{y}$ was replaced by (E.15) and $\boldsymbol{e}$ was assumed to be a white Gaussian noise. Under the assumption that $\boldsymbol{y} \sim \mathcal{CN}(\boldsymbol{y}; \boldsymbol{0}, \sigma^2\boldsymbol{I}_K)$, where σ^2 is the noise variance, the likelihood is approximately given by

$$p(\boldsymbol{x}_1|\boldsymbol{x}_0, \boldsymbol{h}, \sigma^2) \approx (\pi\sigma^2)^{-K} \exp\left\{\frac{-K}{\sigma^2}S(\boldsymbol{h})\right\} \tag{E.31}$$

where $\boldsymbol{x}_0 = \begin{bmatrix} x(1) & \cdots & x(m-1) \end{bmatrix}^T$ and

$$S(\boldsymbol{h}) \triangleq \boldsymbol{h}^H\hat{\boldsymbol{C}}_f\boldsymbol{h} = K^{-1}(\boldsymbol{x}_1 - \boldsymbol{F}_m\tilde{\boldsymbol{h}})^H(\boldsymbol{x}_1 - \boldsymbol{F}_m\tilde{\boldsymbol{h}}) . \tag{E.32}$$

For non-informative and flat priors, the MAP estimates of the noise variance and the filter coefficient are equal to the maximum likelihood estimates which are

$$\hat{\sigma}^2 = S(\hat{\boldsymbol{h}}) = (\boldsymbol{a}^H \hat{\boldsymbol{C}}_{\mathrm{f}}^{-1} \boldsymbol{a})^{-1} \tag{E.33}$$

$$\hat{\boldsymbol{h}} = \boldsymbol{h}^{\mathrm{Capon}} = \hat{\sigma}^2 \hat{\boldsymbol{C}}_{\mathrm{f}}^{-1} \boldsymbol{a} \ . \tag{E.34}$$

The determinant $|\boldsymbol{F}_m^H \boldsymbol{F}_m|$ can be written as

$$|\boldsymbol{F}_m^H \boldsymbol{F}_m| = K^{m-1} \left| \begin{bmatrix} \tilde{\boldsymbol{a}} & -\boldsymbol{I}_{m-1} \end{bmatrix} \hat{\boldsymbol{C}}_{\mathrm{f}}^{*} \begin{bmatrix} \tilde{\boldsymbol{a}}^H \\ -\boldsymbol{I}_{m-1} \end{bmatrix} \right| \ . \tag{E.35}$$

Since the last factor does not grow with K, the asymptotic approximation is $|\boldsymbol{F}_m^H \boldsymbol{F}_m| \approx K^{m-1}$. Thus, (E.25) gives

$$p(m|\boldsymbol{x}) \propto S(\hat{\boldsymbol{h}})^{-K} K^{-(m-1)} \tag{E.36}$$

which is the same as the Bayesian information criterion (BIC) or Schwarz criterion [10].

We use the approximate expression for $p(m|\boldsymbol{x})$ to compute the amplitude estimate averaged over all models. We call this estimator for the model averaged amplitude spectral Capon (MAASC) estimator, and it is given by

$$\hat{\alpha} = \sum_{m=1}^{\lfloor N/2 \rfloor} p(m|\boldsymbol{x}) E\{p(\alpha|\boldsymbol{x}, m)\} = \sum_{m=1}^{\lfloor N/2 \rfloor} p(m|\boldsymbol{x}) \hat{\alpha}_m \ . \tag{E.37}$$

Thus, the MAASC estimate is a weighted sum of the ASC estimates for each filter lengths with the weight determined by the probability of each filter length. As with the ASC estimate, the MAASC estimate can be computed using either the forward or the forward-backward covariance matrix estimate.

4 Iterative Computation of the Inverse Covariance Matrix

The major contribution to the computational complexity in the ASC estimator is the inversion of the covariance matrix estimate. For the MAASC estimator this contribution is even more pronounced as we have to do the inversion for every filter length. In this section, we derive an iterative algorithm for computing the inverse of the forward estimate of the covariance matrix.

For any filter length m, it follows from (E.13) that the forward estimate of the covariance matrix scaled by a factor of K is $\boldsymbol{X}_m \boldsymbol{X}_m^H$ where $\boldsymbol{X}_m$ is defined in (E.12). This scaled estimate is related to $\boldsymbol{X}_{m+1} \boldsymbol{X}_{m+1}^H$ by

$$\boldsymbol{X}_{m+1} \boldsymbol{X}_{m+1}^H = \begin{bmatrix} q_{m+1} & \boldsymbol{r}_{m+1}^H \\ \boldsymbol{r}_{m+1} & \boldsymbol{D}_{m+1} \end{bmatrix} \tag{E.38}$$

where we have defined

$$q_{m+1} \triangleq \boldsymbol{x}_1^H \boldsymbol{x}_1 - |x(m)|^2 \tag{E.39}$$

$$\boldsymbol{r}_{m+1} \triangleq \begin{bmatrix} \boldsymbol{0} & \boldsymbol{X}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1^* \\ 0 \end{bmatrix} \tag{E.40}$$

$$\boldsymbol{D}_{m+1} \triangleq \boldsymbol{X}_m \boldsymbol{X}_m^H - \boldsymbol{x}(N)\boldsymbol{x}^H(N) \ . \tag{E.41}$$

Using blockwise matrix inversion, $(\boldsymbol{X}_{m+1}\boldsymbol{X}_{m+1}^H)^{-1}$ is

$$(\boldsymbol{X}_{m+1}\boldsymbol{X}_{m+1}^H)^{-1} = \begin{bmatrix} \lambda_{m+1} & \boldsymbol{\phi}_{m+1}^H \\ \boldsymbol{\phi}_{m+1} & \boldsymbol{\Psi}_{m+1} \end{bmatrix} \tag{E.42}$$

where we have defined

$$\lambda_{m+1} \triangleq (q_{m+1} - \boldsymbol{r}_{m+1}^H \boldsymbol{D}_{m+1}^{-1} \boldsymbol{r}_{m+1})^{-1} \tag{E.43}$$

$$\boldsymbol{\phi}_{m+1} \triangleq -\lambda_{m+1} \boldsymbol{D}_{m+1}^{-1} \boldsymbol{r}_{m+1} \tag{E.44}$$

$$\boldsymbol{\Psi}_{m+1} \triangleq \boldsymbol{D}_{m+1}^{-1} + \lambda_{m+1}^{-1} \boldsymbol{\phi}_{m+1} \boldsymbol{\phi}_{m+1}^H \ . \tag{E.45}$$

The inverse of $\boldsymbol{X}_{m+1}\boldsymbol{X}_{m+1}^H$ is related the inverse of $\boldsymbol{X}_m\boldsymbol{X}_m^H$ through $\boldsymbol{D}_{m+1}^{-1}$ which can be written as

$$\boldsymbol{D}_{m+1}^{-1} = [\boldsymbol{X}_m \boldsymbol{X}_m^H - \boldsymbol{x}(N)\boldsymbol{x}^H(N)]^{-1} \tag{E.46}$$

$$= (\boldsymbol{X}_m \boldsymbol{X}_m^H)^{-1} + \frac{(\boldsymbol{X}_m \boldsymbol{X}_m^H)^{-1} \boldsymbol{x}(N)\boldsymbol{x}^H(N)(\boldsymbol{X}_m \boldsymbol{X}_m^H)^{-1}}{1 - \boldsymbol{x}^H(N)(\boldsymbol{X}_m \boldsymbol{X}_m^H)^{-1}\boldsymbol{x}(N)} \tag{E.47}$$

by using the matrix inversion lemma. Thus, we can iteratively compute the inverse of the forward estimate of the covariance matrices for all filter lengths without doing any matrix inversions. Whether a similar algorithm for the FB estimate of the covariance matrix exists or not is still an open issue.

5 Simulations

We evaluate the f-MAASC and the fb-MAASC estimators on the same synthetic signal as used in [5]. This signal is the sum of 13 sinusoids at the angular frequencies $2\pi(0.0625, 0.0875, 0.25, 0.285, 0.33, 0.35, 0.37, 0.39, 0.41, 0.43, 0.45, 0.47, 0.49)$, and these frequencies are marked with at dashed line in Fig. E.1. All the sinusoidal components have a phase of $\pi/4$, and the amplitudes of the sinusoids are 1 for the first three, 0.3 for the fourth, and 0.1 for the rest. Our data set consists of $N = 64$ observations from a noise corrupted version of the sinusoidal signal where the noise is complex, white,

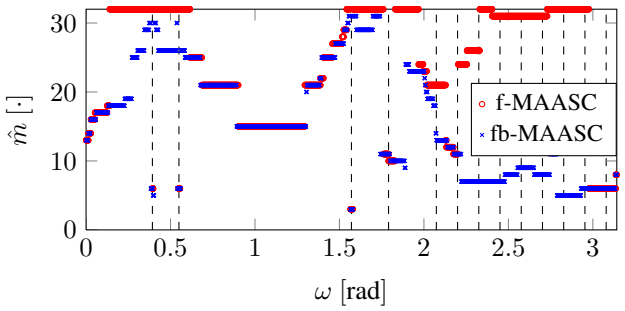

Fig. E.1: The optimal filter length as a function of the frequency.

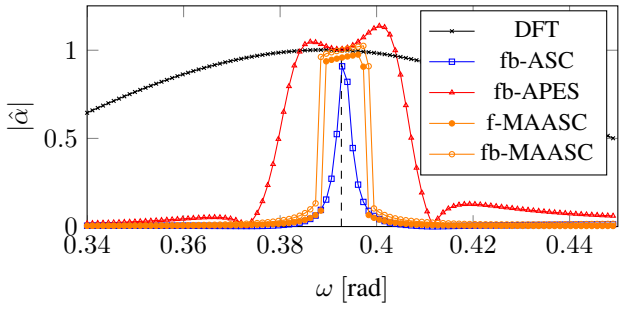

Fig. E.2: The amplitude estimate for various estimators. The filter length for the fb-ASC and the fb-APES was $K = 32$.

and Gaussian with variance σ^2. As in [5], we define the signal-to-noise ratio (SNR) of the signal as the local SNR of the first sinusoid

$$\eta \triangleq 10 \log_{10}(|\alpha_1|^2 \sigma^{-2}) \ . \tag{E.48}$$

Fig. E.1 shows the MAP estimate $\hat{m}$ of the model index as a function of the frequency ω. We see that $\hat{m}$ is large when ω is close to one of the sinusoidal components whereas $\hat{m}$ is small when ω is either at a sinusoidal component or far away from one of the sinusoidal components. This is reasonable as a very long filter is necessary to filter out a sinusoidal component close to ω. On the other hand, a short filter length at a sinusoidal component means that the output vector $\boldsymbol{y}$ has more elements so that we may estimate the complex amplitude with a higher accuracy.

In Fig. E.2, an example of the estimated amplitude spectrum is shown at the first sinusoidal component for an SNR of 20 dB. Despite the fact that the f-MAASC and fb-MAASC estimators are model averaged estimators, their resolution is only slightly worse than the resolution of the fb-ASC estimator with a maximum filter length of

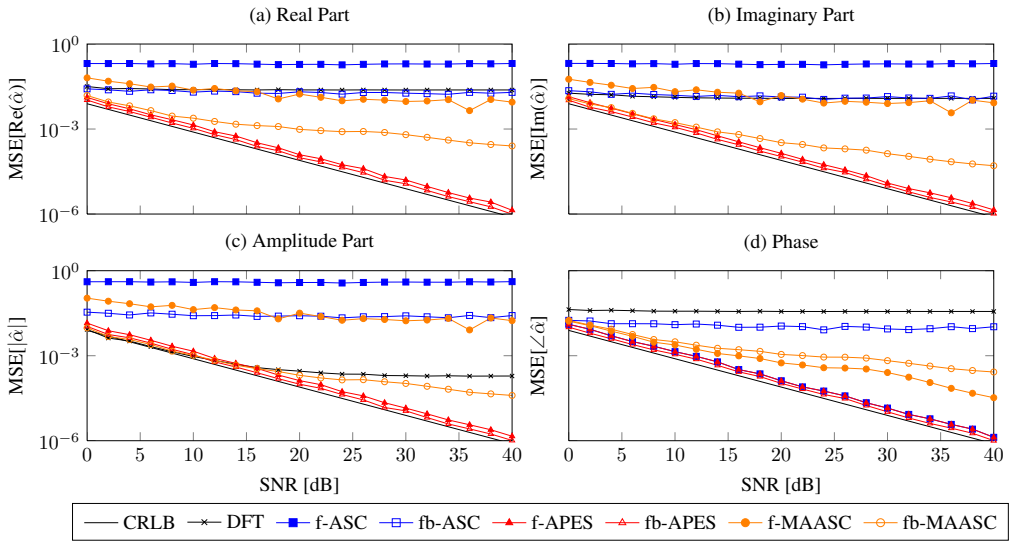

Fig. E.3: The mean squared error (MSE) for various estimators based on 500 Monte Carlo iterations. The filter length for the f-ASC, the fb-ASC, the f-APES, and the fb-APES was $K = 32$.

$K = 32$. However, the resolution was better than for the fb-APES estimator.

Based on 500 Monte Carlo iterations, the mean squared errors (MSE) at the first sinusoidal component for the estimates of the real and complex part of the complex amplitude, the amplitude, and the phase are shown and compared to the asymptotic Cramer-Rao lower bound (CRLB) in Fig. E.3. Clearly, the f-MAASC and fb-MAASC estimators significantly reduce the MSE of the corresponding f-ASC and fb-ASC estimators. Other simulations have shown that this reduction comes from a reduction in both the bias and the variance. Although the maximum filter length of $K = 32$ is not a recommended length of the ASC-filters (see Sec. 1), we used this length in our simulations to see how much the bias was lowered by the MAASC-estimators while still having nearly the same resolution as demonstrated in Fig. E.2. If the filter length is reduced, the MSE-performance of the ASC-filters also improves significantly, but at the expense of a coarser resolution.

Note that the f-APES and the fb-APES can also be cast in a model averaged framework. However, we observed only minor improvements in the MSE at the cost of a slightly worse resolution and a higher computational complexity. On the other hand, the model averaging attenuates the line-splitting (see Fig. E.2) in the APES estimate of the amplitude spectrum.

6 Conclusion

In this paper, we have proposed a data-adaptive way of determining the filter length of the Capon filter. The adaptation was based on the approximate MAP approach by Djuric, and it led to a very simple way of computing an approximate posterior distribution for the filter length. Based on this posterior distribution, we also derived a model averaged amplitude spectral Capon estimator for both the forward and the forward-backward estimate of the covariance matrix. For nearly the same resolution, simulations on a synthetic signal showed that the f-MAASC and fb-MAASC significantly lowered the mean squared error of the complex amplitude estimates as compared to the traditional forward and forward-backward amplitude spectral Capon estimators, respectively.

References

[1] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Prentice Hall, May 2005.

[2] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, B. H. Juang, Ed. Morgan & Claypool, 2009.

[3] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[4] J. Li and P. Stoica, "An adaptive filtering approach to spectral estimation and SAR imaging," *IEEE Trans. Signal Process.*, vol. 44, no. 6, pp. 1469–1484, Jun. 1996.

[5] P. Stoica, A. Jakobsson, and J. Li, "Matched-filter bank interpretation of some spectral estimators," *Elsevier Signal Processing*, vol. 66, no. 1, pp. 45–59, 1998.

[6] H. Li, J. Li, and P. Stoica, "Performance analysis of forward-backward matched-filterbank spectral estimators," *IEEE Trans. Signal Process.*, vol. 46, no. 7, pp. 1954–1966, Jul. 1998.

[7] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results, and an application," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 338–352, Feb. 2000.

[8] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726–2735, Oct. 1998.

[9] A. Jakobsson and P. Stoica, "Combining Capon and APES for estimation of spectral lines," *Circ. Syst. Signal Process.*, vol. 19, no. 2, pp. 159–169, 2000.

[10] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

Paper F

On Compressed Sensing and the Estimation of Continuous Parameters From Noisy Observations

Jesper Kjær Nielsen, Mads Græsbøll Christensen, and Søren Holdt Jensen

The paper has been published in the
Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, Mar. 2012.

Abstract

Compressed sensing (CS) has in recent years become a very popular way of sampling sparse signals. This sparsity is measured with respect to some known dictionary consisting of a finite number of atoms. Most models for real world signals, however, are parametrised by continuous parameters corresponding to a dictionary with an infinite number of atoms. Examples of such parameters are the temporal and spatial frequency. In this paper, we analyse how CS affects the estimation performance of any unbiased estimator when we assume such infinite dictionaries. We base our analysis on the Cramer-Rao lower bound (CRLB) which is frequently used for benchmarking the estimation accuracy of unbiased estimators. For the popular sensing matrices such as the Gaussian sensing matrix, our analysis shows that compressed sensing on average degrades the estimation accuracy by at least the down-sample factor.

1 Introduction

For a wide range of applications such as compression, enhancement, identification, and separation, sparse decompositions have been a very useful tool. Mathematically speaking, a sparse decomposition of an N-dimensional complex vector $\boldsymbol{x} \in \mathbb{C}^N$ can be written as the linear model

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{s} + \boldsymbol{e} \tag{F.1}$$

where $\boldsymbol{\Psi} \in \mathbb{C}^{N \times D}$ is referred to as the basis or dictionary, $\boldsymbol{s}$ is a D-dimensional S-sparse vector, and $\boldsymbol{e} \in \mathbb{C}^N$ is an error vector modelling noise and model inaccuracies. By S-sparse, we mean that $\boldsymbol{s}$ contains exactly S non-zero coefficients and $D - S$ zeros. Moreover, we say that $\boldsymbol{x}$ is an S-sparse or a compressible signal in the basis $\boldsymbol{\Psi}$ if $\boldsymbol{e} = 0$ or $\boldsymbol{e} \approx 0$, respectively. Traditionally, the non-zero coefficients of $\boldsymbol{s}$ are found by greedy or ℓ_1 optimisation algorithms such as matching pursuit [1] or basis pursuit [2]. In the usual cases where $N \gg S$, however, these methods may suffer from a large computational overhead as they have to work directly on $\boldsymbol{x}$. In compressed sensing (CS) [3], this overhead is decreased considerably by utilising the sparsity during the data acquisition. That is, instead of acquiring $\boldsymbol{x}$ by sampling at the Nyquist rate, we acquire $\boldsymbol{y} \in \mathbb{C}^M$ with $S < M \leq N$ by only collecting an amount of data close to the sparsity level S. Thus, CS may enable a faster computation of the parameters, data acquisition at a lower sample rate, and less demanding storage requirements. These properties are very important for most signal processing algorithms, and CS has therefore become very popular. Mathematically speaking, we model the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ by $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$ where $\boldsymbol{\Phi} \in \mathbb{C}^{M \times N}$ is referred to as the sensing matrix. If $\boldsymbol{x}$ is compressible, and $\boldsymbol{\Phi}$ is chosen appropriately, the vector $\boldsymbol{s}$ can be computed directly from $\boldsymbol{y}$ provided that the Restricted Isometry Property (RIP) holds [3, 4]. Until recently [5], the recovery was only shown to hold for the orthogonal or incoherent dictionaries $\boldsymbol{\Psi}$. Consequently, much

attention has been directed towards finding sensing matrices which make M as small as possible given an incoherent dictionary [6, 7]. The dictionary $\boldsymbol{\Psi}$ in (F.1) consists of D column vectors $\{\boldsymbol{\psi}_d\}_{d=1}^{D}$ which are often referred to as atoms, and popular choices of the dictionary are the Fourier basis and a wavelet basis. Typically, the atoms can be represented by a parametric function $\boldsymbol{f}(\boldsymbol{\phi})$, and each atom is constructed by selecting a specific value $\boldsymbol{\psi}_d = \boldsymbol{f}(\boldsymbol{\phi}_d)$ for the parameter of this function. For example, the atoms of the incoherent Fourier basis is formed by sampling the frequency parameter $\phi = \omega$ on the Fourier grid $\phi_d = 2\pi(d-1)/D$ with $D = N$. For most real world signals, however, the parameter $\boldsymbol{\phi}$ is a continuous parameter corresponding to highly coherent dictionaries with $D \to \infty$. A sparse decomposition with a finite dictionary is therefore in direct contradiction with the physics behind most signal models of the form

$$\boldsymbol{x} = \boldsymbol{A}(\boldsymbol{\phi})\boldsymbol{\alpha} + \boldsymbol{w} \tag{F.2}$$

where $\boldsymbol{w} \in \mathbb{C}^N$ is a noise vector, and $\boldsymbol{A}(\boldsymbol{\phi}) \in \mathbb{C}^{N\times S}$ is parametrised by $\boldsymbol{\phi} \in \mathbb{C}^{K-S}$ and contains the S true atoms with the amplitudes $\boldsymbol{\alpha} \in \mathbb{C}^S$. The scalar K is the total number of variables in $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$. Comparing the models in (F.2) and (F.1), we see that $\boldsymbol{A}(\boldsymbol{\phi})\boldsymbol{\alpha} \approx \boldsymbol{\Psi}\boldsymbol{s}$ with equality if the true atoms in $\boldsymbol{A}(\boldsymbol{\phi})$ are included in $\boldsymbol{\Psi}$. As demonstrated in [8], we obtain an inferior compression scheme by using the model in (F.1) rather than (F.2) when equality does not hold. CS has been developed under the assumption that equality holds. In other words, $\boldsymbol{\phi}$ is assumed to be a discrete parameter whose possible values are used to construct the atoms of the dictionary. When CS is viewed in this light, we may interpret the RIP as a requirement to the distance between adjacent values that $\boldsymbol{\phi}$ may take.

In this paper, we do not assume that $\boldsymbol{\phi}$ is a discrete parameter. For various popular sensing matrices [4, 9], we instead investigate the accuracy with which we can estimate the continuous parameters of the model in (F.2) when we are giving $\boldsymbol{y}$ instead of $\boldsymbol{x}$. For the MUSIC algorithm, we noted a significant loss in the estimation accuracy in [10]. Here, however, we do not consider a specific estimation algorithm, but only the best possible performance that any unbiased estimation algorithm can obtain. We therefore base the analysis on the Cramer-Rao lower bound (CRLB) which has previously [11, 12] been used to assess the estimation accuracy of the non-zero elements of $\boldsymbol{s}$, assuming a finite dictionary. In this paper, however, we work directly with the model in (F.2), corresponding to an infinite dictionary. The paper is organised as follows: In Sec. 2, we present the CRLB for the model in (F.2). The CRLB is used to benchmark the performance of unbiased estimators, and we modify it to the situation in which CS is used in Sec. 3. In Sec. 4, we establish a connection between the CRLB with and without CS by deriving a lower bound on the expected CRLB for some of the popular sensing matrices. An illustrative simulation is presented in Sec. 5, and Sec. 6 concludes this paper.

2 Cramer-Rao Lower Bound

Consider the general problem in which we observe N random data points $\boldsymbol{x}$ which we mathematically describe by a family of probability density functions $p(\boldsymbol{x};\boldsymbol{\theta})$. Without loss of generality, we assume that this model is parametrised by the real[1] parameter vector $\boldsymbol{\theta}$ which we wish to estimate based on the data. In order to do this, we construct an unbiased estimator $\hat{\boldsymbol{\theta}}$ which maps the data into an estimate. For the covariance matrix $\boldsymbol{C}_{\hat{\boldsymbol{\theta}}}$ of any unbiased estimator, the CRLB guarantees that $\boldsymbol{C}_{\hat{\boldsymbol{\theta}}} - \boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}) \geq \boldsymbol{0}$ where the inequality denotes positive semi-definitiveness. Thus, for the variance of the estimator for the k'th parameter, we have that

$$\mathrm{var}(\hat{\theta}_k) = \left[\boldsymbol{C}_{\hat{\boldsymbol{\theta}}}\right]_{kk} \geq \left[\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta})\right]_{kk} \tag{F.3}$$

where $[\cdot]_{kk}$ denotes the (k,k)'th element. The matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ is the Fisher information matrix (FIM), and it is given by [13]

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = E\left\{\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^T\right\} \tag{F.4}$$

where $(\cdot)^T$ denotes matrix transpose. It can be shown that if $\boldsymbol{x}$ has a multivariate complex normal distribution whose mean and covariance are parametrised by $\boldsymbol{\theta}$, i.e., $\boldsymbol{x} \sim \mathcal{CN}(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{C}(\boldsymbol{\theta}))$, then the (k,l)'th element of the FIM is given by [13]

$$\begin{aligned}\left[\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\right]_{kl} = 2\mathrm{Re}&\left[\frac{\partial \boldsymbol{\mu}^H(\boldsymbol{\theta})}{\partial \theta_k}\boldsymbol{C}^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_l}\right] \\ &+ \mathrm{tr}\left[\boldsymbol{C}^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{C}(\boldsymbol{\theta})}{\partial \theta_k}\boldsymbol{C}^{-1}(\boldsymbol{\theta})\frac{\partial \boldsymbol{C}(\boldsymbol{\theta})}{\partial \theta_l}\right] .\end{aligned} \tag{F.5}$$

where $(\cdot)^H$ denotes complex transposition, $\mathrm{tr}(\cdot)$ denotes matrix trace, and $\mathrm{Re}[\cdot]$ takes the real part of a complex number. For the model in (F.2) with $\boldsymbol{w} \sim \mathcal{CN}(\boldsymbol{0}, \sigma_w^2\boldsymbol{I}_N)$, the $(K+1)$-dimensional parameter vector is $\boldsymbol{\theta} \triangleq \begin{bmatrix}\boldsymbol{\phi}^T & \boldsymbol{\alpha}^T & \sigma_w^2\end{bmatrix}^T$, and we have that

$$\boldsymbol{x} \sim \mathcal{CN}(\boldsymbol{A}(\boldsymbol{\phi})\boldsymbol{\alpha}, \sigma_w^2\boldsymbol{I}_N) \tag{F.6}$$

where $\boldsymbol{I}_N$ is the N-dimensional identity matrix. Using (F.5) and (F.6), the FIM is given by

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = \begin{bmatrix}2\sigma_w^{-2}\mathrm{Re}(\boldsymbol{Q}^H\boldsymbol{Q}) & \boldsymbol{0} \\ \boldsymbol{0}^T & N\sigma_w^{-4}\end{bmatrix} \tag{F.7}$$

[1] If the model is parametrised by complex parameters $\boldsymbol{\xi} = \boldsymbol{\xi}_r + j\boldsymbol{\xi}_i$, say, then $\boldsymbol{\theta}$ is defined as $\boldsymbol{\theta} \triangleq \begin{bmatrix}\boldsymbol{\xi}_r^T & \boldsymbol{\xi}_i^T\end{bmatrix}^T$.

where we have defined

$$\boldsymbol{q}_k \triangleq \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_k} \qquad \text{for } k = 1, 2, \ldots, K \tag{F.8}$$

$$\boldsymbol{Q} \triangleq \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 & \cdots & \boldsymbol{q}_K \end{bmatrix} . \tag{F.9}$$

3 The Expected Projection Matrix

As mentioned in the introduction, we observe $\boldsymbol{y}$ instead of $\boldsymbol{x}$ in compressed sensing. Incorporating this into (F.2) yields

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{\Phi}\boldsymbol{A}(\boldsymbol{\phi})\boldsymbol{\alpha} + \boldsymbol{\Phi}\boldsymbol{w} \tag{F.10}$$

from which we see that

$$\boldsymbol{y}|\boldsymbol{\Phi} \sim \mathcal{CN}(\boldsymbol{\Phi}\boldsymbol{A}(\boldsymbol{\phi})\boldsymbol{\alpha}, \sigma_w^2 \boldsymbol{\Phi}\boldsymbol{\Phi}^H) . \tag{F.11}$$

In the following sections, we investigate how the sensing matrix affects the CRLB derived in Sec. 2.

3.1 Fisher Information Matrix in Compressed Sensing

For $k, l \leq K$, we obtain from (F.5) and (F.11) that

$$\begin{aligned} [\boldsymbol{\mathcal{I}}_{\text{CS}}(\boldsymbol{\theta})]_{kl} &= 2\sigma_w^{-2} \text{Re}\left[\boldsymbol{q}_k^H \boldsymbol{\Phi}^H (\boldsymbol{\Phi}\boldsymbol{\Phi}^H)^{-1} \boldsymbol{\Phi} \boldsymbol{q}_l\right] \\ &= 2\sigma_w^{-2} \text{Re}\left[\boldsymbol{q}_k^H \boldsymbol{\Pi} \boldsymbol{q}_l\right] \end{aligned} \tag{F.12}$$

where we have defined $\boldsymbol{\Pi} \triangleq \boldsymbol{\Phi}^H (\boldsymbol{\Phi}\boldsymbol{\Phi}^H)^{-1}\boldsymbol{\Phi}$ which is an $N \times N$ orthogonal projection matrix of rank M. For $k = l = K + 1$, we can rewrite (F.5) as

$$\begin{aligned} [\boldsymbol{\mathcal{I}}_{\text{CS}}(\boldsymbol{\theta})]_{kl} &= \text{tr}\left\{\sigma_w^{-2}(\boldsymbol{\Phi}\boldsymbol{\Phi}^H)^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^H \sigma_w^{-2}(\boldsymbol{\Phi}\boldsymbol{\Phi}^H)^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^H\right\} \\ &= \text{tr}\left\{\sigma_w^{-4}\boldsymbol{I}_M\right\} = M\sigma_w^{-4} . \end{aligned} \tag{F.13}$$

Thus, the FIM is given by

$$\boldsymbol{\mathcal{I}}_{\text{CS}}(\boldsymbol{\theta}) = \begin{bmatrix} 2\sigma_w^{-2}\text{Re}(\boldsymbol{Q}^H\boldsymbol{\Pi}\boldsymbol{Q}) & \boldsymbol{0} \\ \boldsymbol{0}^T & M\sigma_w^{-4} \end{bmatrix} . \tag{F.14}$$

Compared against the FIM in (F.7), we see that $\boldsymbol{\mathcal{I}}_{\text{CS}}(\boldsymbol{\theta})$ differs in terms of the scaling factor of the $(K+1, K+1)$'th element and the inclusion of the projection matrix $\boldsymbol{\Pi}$ inside the inner matrix product $\boldsymbol{Q}^H\boldsymbol{Q}$. The interpretation of the latter is straightforward; we project the columns of $\boldsymbol{Q}$ onto the subspace spanned by $\boldsymbol{\Phi}^H$. Therefore, the diagonal elements of $\boldsymbol{\mathcal{I}}_{\text{CS}}(\boldsymbol{\theta})$ decrease compared against the corresponding elements of $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ unless $\boldsymbol{Q}$ is spanned by $\boldsymbol{\Phi}^H$.

3.2 Typical Sensing Matrices

As alluded to in the introduction, the choice of sensing matrix $\mathbf{\Phi}$ is vital in CS; we wish to find a sensing matrix that obeys the RIP for as few measurements M as possible. Perhaps surprisingly, stochastic sensing matrices have been shown to be nearly optimal for almost any choice of basis $\mathbf{\Psi}$ [14]. Some of the most popular choices are listed below [4].

1. Select the entries of $\mathbf{\Phi}$ as i.i.d. samples from a normal distribution with variance $1/M$.
2. Sample N N-dimensional i.i.d. vectors from a normal distribution with unit variance. Find an orthonormal basis of these N random vectors and select the rows of $\mathbf{\Phi}$ as M random rows from this orthonormal basis.
3. Sample N M-dimensional i.i.d. vectors uniformly at random from the unit sphere.
4. Select the entries of $\mathbf{\Phi}$ as i.i.d. samples from a symmetric Bernoulli distribution with outcomes $\pm 1/M$.

Once the sensing matrix has been selected, the FIM is easy to calculate. Since the sensing matrix is often selected at random, however, it is not particularly useful to say something about a specific realisation of the sensing matrix. Therefore, the next logical step in our analysis is to investigate the statistics of the inverse FIM when the sensing matrix is selected at random from some matrix variate distribution. Unfortunately, this is in general a very hard problem, and we therefore consider the simpler task of investigating the expected FIM for the various sensing matrices. We use this to derive a lower bound on the expected inverse FIM in Sec. 4.

Since the unknown parameters $\boldsymbol{\theta}$ are assumed to be deterministic variables, it readily follows from (F.14) that the expected FIM is given by

$$E\left\{\mathcal{I}_{\mathrm{CS}}(\boldsymbol{\theta})\right\} = \begin{bmatrix} 2\sigma_w^{-2}\mathrm{Re}[\boldsymbol{Q}^H E\{\mathbf{\Pi}\}\boldsymbol{Q}] & \mathbf{0} \\ \mathbf{0}^T & M\sigma_w^{-4} \end{bmatrix}. \tag{F.15}$$

Thus, in order to find the expected FIM, we have to find the expected projection matrix $E\{\mathbf{\Pi}\} = E\{\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}\}$. For this purpose, we use the following theorem.

Theorem 1
Let $\mathbf{\Phi}$ be a random $M \times N$ matrix with $M < N$, rank M almost everywhere, and the probability density function (pdf) $f_{\mathbf{\Phi}}(\mathbf{\Phi})$. Furthermore, let $\mathbf{\Pi} = \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}$ be the $N \times N$ orthogonal projection matrix of rank M onto the subspace spanned by the rows of $\mathbf{\Phi}$ and denote the space of points corresponding to all such projection matrices by $P_{M,N-M}$. If $f_{\mathbf{\Phi}}(\mathbf{\Phi})$ is invariant under the right-orthogonal transformation $\mathbf{\Phi} \to \mathbf{\Phi}\boldsymbol{R}$ for any $N \times N$ orthogonal matrix $\boldsymbol{R}$, then the PDF of $\mathbf{\Pi}$ is uniform on $P_{M,N-M}$ and $E\{\mathbf{\Pi}\} = (M/N)\boldsymbol{I}_N$.

Sketch of the Proof. It follows from [15, Th. 2.4.3] that $\boldsymbol{H} = (\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1/2}\boldsymbol{\Phi}$ is uniformly distributed on the Stiefel manifold $V_{M,N}$ if $f_{\boldsymbol{\Phi}}(\boldsymbol{\Phi})$ is invariant under any right-orthogonal transformation. It then follows from [15, Th. 2.2.1] that $\boldsymbol{Q}\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{Q}^T$ is uniformly distributed on $P_{M,N-M}$ for any orthogonal $\boldsymbol{Q}$ independent of $\boldsymbol{H}^T\boldsymbol{H}$. Now, since $\boldsymbol{\Pi} = \boldsymbol{H}^T\boldsymbol{H}$, it therefore follows that $\boldsymbol{\Pi}$ is uniformly distributed on $P_{M,N-M}$. Moreover, if follows that

$$E\{\boldsymbol{\Pi}\} = \boldsymbol{Q}E\{\boldsymbol{\Pi}\}\boldsymbol{Q}^T \implies E\{\boldsymbol{\Pi}\} \propto \boldsymbol{I}_N \tag{F.16}$$

$$\mathrm{tr}(E\{\boldsymbol{\Pi}\}) = E\{\mathrm{tr}(\boldsymbol{\Pi})\} = M \tag{F.17}$$

Eq. (F.17) follows from the fact that a projection matrix of rank M has exactly M ones and $N-M$ zeros as its eigenvalues. Combining (F.16) and (F.17) readily gives $E\{\boldsymbol{\Pi}\} = (M/N)\boldsymbol{I}_N$. □

By use of Theorem 1, we can show that the sensing matrices of type 1 and 2 from above result in an expected projection matrix of $E\{\boldsymbol{\Pi}\} = (M/N)\boldsymbol{I}_N$. As shown in Sec. 5, empirical evidence also suggests that this is the case for the sensing matrices of type 3 and 4. Interestingly, for the simplest possible sensing matrix, the Kronecker sensing matrix which is the identity matrix with $N-M$ random rows removed, we obtain the same expected projection matrix. To see this, consider that the projection matrix corresponding to such a Kronecker sensing matrix $\boldsymbol{K}$ is given by $\boldsymbol{\Pi} = \boldsymbol{K}^T\boldsymbol{K}$ which is a diagonal matrix with M ones and $N-M$ zeros uniformly distributed on the diagonal. Thus, there are

$${}_N C_M = \binom{N}{M} = \frac{N!}{M!(N-M)!} \tag{F.18}$$

distinct projection matrices each with probability $1/{}_N C_M$. The expected value of the projection matrix is therefore

$$E\{\boldsymbol{\Pi}\} = \frac{1}{{}_N C_M}\sum_{i=1}^{{}_N C_M}\boldsymbol{\Pi}_i = \frac{{}_{N-1}C_{M-1}}{{}_N C_M}\boldsymbol{I}_N = \frac{M}{N}\boldsymbol{I}_N \tag{F.19}$$

where the second equality follows from the fact that there are ${}_{N-1}C_{M-1}$ projection matrices $\boldsymbol{\Pi}$ with a one on the k'th diagonal element.

4 A Bound on the Expected CRLB

Inserting the expected projection matrix of $E\{\boldsymbol{\Pi}\} = (M/N)\boldsymbol{I}_N$ into (F.15) yields

$$E\left\{\boldsymbol{\mathcal{I}}_{\mathrm{CS}}(\boldsymbol{\theta})\right\} = \frac{M}{N}\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})\ . \tag{F.20}$$

Thus, on average, the elements of the FIM with CS is M/N times the elements of the FIM without CS. The expected inverse FIM can be bounded as

$$E\left\{\boldsymbol{\mathcal{I}}_{\mathrm{CS}}^{-1}(\boldsymbol{\theta})\right\} \geq \left(E\left\{\boldsymbol{\mathcal{I}}_{\mathrm{CS}}(\boldsymbol{\theta})\right\}\right)^{-1} = \frac{N}{M}\boldsymbol{\mathcal{I}}^{-1}(\boldsymbol{\theta}) \ . \tag{F.21}$$

Since $M < N$ in CS, the expected CRLB with CS increases as compared to the case without CS. Interestingly, the popular Gaussian and nearly orthogonal sensing matrices do not on average perform better than the Kronecker sensing matrix from an estimation theoretic point of view. Thus, on average, the Gaussian and the nearly orthogonal sensing matrices of the form outlined above decrease the estimation accuracy by an amount equal to the case where random samples are simply thrown away. Furthermore, the expected estimation accuracy decreases inversely proportional to at least the number of samples that we retain in the data acquisition step of CS. This was also demonstrated in [10] in which the direction-of-arrival estimation accuracy for a varying M was compared to the CRLB. Can we do any better than this? That is, can we select a sensing matrix such that the elements of the inverse FIM with CS are closer to the elements of the CRLB without CS? To answer this question, we first take a closer look at the orthogonal projection matrix $\boldsymbol{\Pi}$. From (F.17), we have that the trace of an expected projection matrix must be equal to M. All projection matrices must fulfil this constraint so if we wish to construct a diagonal expected projection matrix with equal elements, we therefore have that $E\{\boldsymbol{\Pi}\} = (M/N)\boldsymbol{I}_N$. This result is the same as the expected projection matrix for the popular sensing matrices presented above. However, as we saw in Sec. 3.1, it is possible to design a sensing matrix such that the CRLB is unaffected provided that the columns of $\boldsymbol{Q}$ are spanned by the rows of $\boldsymbol{\Phi}$. Unfortunately, since the design of such a sensing matrix requires that we know the parameters we wish to estimate, it is infeasible, unless we have a strong prior knowledge about the values of the missing parameters. In this case, however, it may be better to employ Bayesian inference methods which offer a unified way of incorporating prior knowledge.

5 Simulations

We demonstrate the validity of our analysis on a simple but well-known example. In the example, we consider a complex sinusoid in complex white Gaussian noise, i.e.,

$$x_n = \alpha e^{(j\omega n + j\varphi)} + w_n \ , \quad \text{for } n = 0, \ldots, N-1 \tag{F.22}$$

where $\alpha > 0$, $\varphi \in [-\pi, \pi]$ and $\omega \in [-\pi, \pi]$ are the amplitude, phase and (angular) frequency, respectively. The noise variance is σ_w^2. The CRLB for this signal is well-

	E-CRLB-CS	Type 1	Type 2	Type 3	Type 4	Kron
$\mathrm{E}\{\mathrm{var}(\hat{\alpha})\}$	3.125	3.363	3.366	3.365	3.371	3.125
$\mathrm{E}\{\mathrm{var}(\hat{\phi})\}$	12.21	13.42	13.41	13.40	13.41	13.31
$\mathrm{E}\{\mathrm{var}(\hat{\omega})\}$	0.0092	0.0101	0.0101	0.0101	0.0101	0.0101

Table F.1: The mean value for the CRLB based on 100,000 Monte Carlo runs. In each run, sensing matrices of type 1-4 and the Kronecker matrix all of size 16×64 were generated. All values are scaled by a factor of 1000.

known and given by [13]

$$\mathrm{var}(\hat{\alpha}) \geq \frac{\sigma_w^2}{2N} \qquad \mathrm{var}(\hat{\varphi}) \geq \frac{\sigma_w^2(2N-1)}{\alpha^2 N(N+1)}$$
$$\mathrm{var}(\hat{\omega}) \geq \frac{6\sigma_w^2}{\alpha^2 N(N^2-1)} \qquad \mathrm{var}(\hat{\sigma}_w^2) \geq \frac{\sigma_w^4}{N}$$

and the k'th row of $\boldsymbol{Q}$ is

$$[\boldsymbol{Q}]_{k:} = \begin{bmatrix} e^{(j\omega n + j\varphi)} & j\alpha e^{(j\omega n + j\varphi)} & j\alpha n e^{(j\omega n + j\varphi)} \end{bmatrix} . \tag{F.23}$$

For each of the four types of sensing matrices and for the Kronecker matrix, we ran 100,000 Monte Carlo runs in which we calculated the inverse FIM given by the inverse of (F.14). The size[2] of the sensing matrices was 16×64. For the diagonal elements of the inverse FIM corresponding to the amplitude, phase and frequency, we calculated their 500 bins normalised histograms and mean values. Fig. F.1 shows the normalised histograms of the CRLB for the frequency parameter. Clearly, the histograms are almost coinciding with the exception of the histogram corresponding to the Kronecker sensing matrix. Fig. F.1 also shows the CRLB without CS as well as the lower bound for the expected CRLB with CS. Table F.1 lists the mean values of the CRLB with CM corresponding to the amplitude, phase and frequency. The lower bound for the expected CRLB is also listed. Again, we see the same pattern; the mean values for the type 1-4 sensing matrices were more or less the same while the mean value for the Kronecker sensing matrix was slightly different. All values were on or above the lower bound for the expected CRLB.

6 Conclusion

In this paper, we have analysed compressed sensing (CS) from an estimation theoretic point of view by use of the Cramer-Rao lower bound (CRLB). Not surprisingly, our

[2] As a rule of thumb, the value of M should be approximately four times the number of unknown parameters [4].

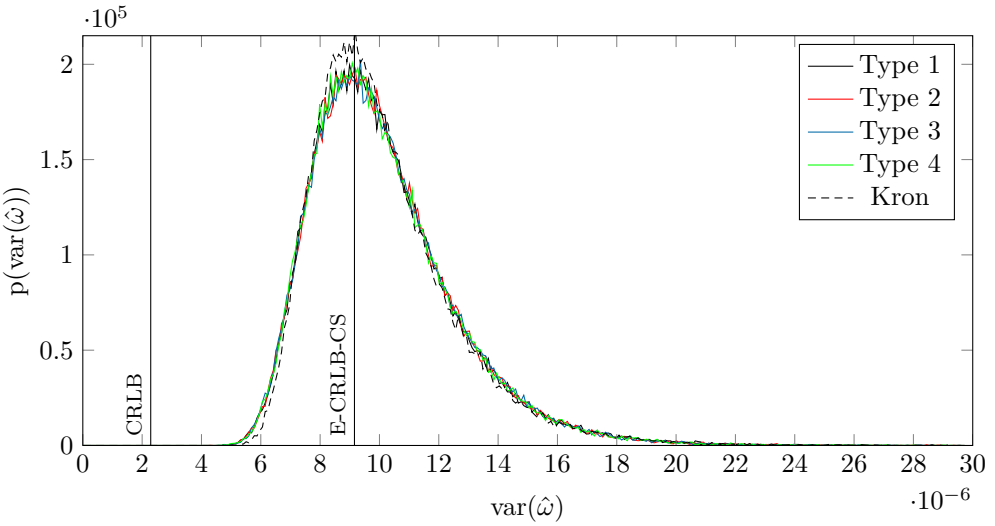

Fig. F.1: Estimated pdf of the CRLB for the frequency parameter based on 100,000 Monte Carlo runs. In each run, sensing matrices of type 1-4 and the Kronecker matrix all of size 16×64 were generated.

analysis have shown that CS on average degrades our ability to estimate continuous parameters. For some of the popular sensing matrices such as the Gaussian sensing matrix, we quantified the expected degradation by showing that the ratio between the expected CRLB with CS and the CRLB without CS is lower bounded by the ratio between the number of columns N and the number of rows M of the sensing matrix. Perhaps more surprisingly, we also showed that the bound is the same for the Kronecker sensing matrix. That is, from an estimation theoretic point of view some of the popular sensing matrices degrade on average our estimation accuracy by an amount equal to the situation in which we throw $N - M$ random samples away.

References

[1] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[2] S. Chen and D. Donoho, "Basis pursuit," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 1994.

[3] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[4] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[5] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 1, pp. 59–73, Jul. 2011.

[6] A. Hormati, A. Karbasi, S. Mohajer, and M. Vetterli, "An estimation theoretic approach for sparsity pattern recovery in the noisy setting," Nov. 2009, unpublished manuscript.

[7] V. Abolghasemi, S. Ferdowsi, B. Makkiabadi, and S. Sanei, "On optimization of the measurement matrix for compressive sensing," in *Proc. European Signal Processing Conf.*, Aug. 2010, pp. 427–431.

[8] M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," 2011, unpulished manuscript.

[9] R. G. Baraniuk, "Compressive sensing," *Lecture Notes in IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118–120, Jul. 2007.

[10] M. G. Christensen and J. K. Nielsen, "Joint direction-of-arrival and order estimation in compressed sensing using angles between subspaces," in *Proc. IEEE Workshop on Stat. Signal Process.*, Jun. 2011, pp. 449–452.

[11] B. Babadi, N. Kalouptsidis, and V. Tarokh, "Asymptotic achievability of the Cramer-Rao bound for noisy compressive sampling," *IEEE Trans. Signal Process.*, vol. 57, no. 3, pp. 1233–1236, Mar. 2009.

[12] Z. Ben-Haim and Y. Eldar, "The Cramer-Rao bound for estimating a sparse parameter vector," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3384–3389, Jun. 2010.

[13] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall PTR, Mar. 1993.

[14] R. G. Baraniuk, M. A. Davenport, R. A. DeVore, and M. B. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, Dec. 2008.

[15] Y. Chikuse, *Statistics on Special Manifolds*, 1st ed. Springer, Feb. 2003.

Paper G

Joint Direction-of-Arrival and Order Estimation in Compressed Sensing using Angles Between Subspaces

Mads Græsbøll Christensen and Jesper Kjær Nielsen

The paper has been published in the
Proceedings of the IEEE Workshop on Statistical Signal Processing, Jun. 2011, pp. 449–452.

The layout has been revised.

Abstract

In this paper, we consider the problem of joint direction-of-arrival and order estimation in array processing with compressed sensing. In particular, we show how to solve these problems jointly using a subspace approach based on the notion of angles between subspaces. In the process, we also discuss the conditions on the measurement matrix and demonstrate how to implement the estimator algorithm efficiently when using compressed sensing. Our simulation results show that it is indeed possible to solve these problems and that good performance can be obtained, although the use of compressed sensing does have an impact on the performance of the estimator.

1 Introduction

A classical problem in array signal processing is that of determining the direction-of-arrival (DOA) of sources impinging on the array, and many methods have been proposed throughout the history, including such prominent work as [1–3]. We are here also interested in this problem, but in a new context, namely that of compressed sensing [4, 5], wherein measurements are formed as random linear combinations of the sensor inputs. In this context, the parameter estimation and signal reconstruction problems are most often dealt with by solving convex problems, typically equality constrained 1-norm minimization problems, and such approaches have also been applied to array signal processing, e.g., [6–8]. In both DOA estimation and compressed sensing, the number of sources, i.e., the model order, is often assumed known, or, is assumed to have been found in some other way. The number of sources may of course vary over time, and the question is, however, how to deal with this both in DOA estimation and in compressed sensing, since the design of an appropriate measurement matrix requires that the number of sources is known, i.e., that the level of spatial sparsity is known a priori.

The problem under consideration can be formally defined as follows. Let $y_k(n)$ be the observed signal at time n for sensor k of a uniform linear array (ULA), and let K be the total number of sensors and N the total number of snapshots. For L narrowband sources (the model order) and complex spatial noise $\mathbf{e}(n)$ impinging on the array, the spatial signal model can be expressed as

$$\mathbf{y}(n) = \mathbf{\Phi}\mathbf{A}\mathbf{x}(n) + \mathbf{\Phi}\mathbf{e}(n) \tag{G.1}$$

where $\mathbf{x}(n)$ contains the individual signals of the sources impinging on the array and $\mathbf{y}(n)$ contains $y_k(n)$ for $k = 1, \ldots, M$. The noise is here assumed to be colored, i.e., $\mathrm{E}\left\{\mathbf{e}(n)\mathbf{e}^H(n)\right\} = \mathbf{Q}$ with $\mathbf{Q}$ being the noise covariance matrix, $\mathrm{E}\{\cdot\}$ the expectation operator and $(\cdot)^H$ the conjugate transpose. We here assume that $\mathbf{Q}$ is known or estimated in some other way and that it is invertible. Moreover, $\mathbf{A} \in \mathbb{C}^{K\times L}$ is a Vandermonde matrix containing the steering vectors of the $L < K$ incoherent sources,

with unknown spatial frequencies $\{\omega_l\}$, defined as $\mathbf{A} = \left[\begin{array}{ccc} \mathbf{a}(\omega_1) & \cdots & \mathbf{a}(\omega_L) \end{array} \right]$ where $\mathbf{a}(\omega_l) = \left[\begin{array}{cccc} 1 & e^{-j\omega_l} & \cdots & e^{-j\omega_l(K-1)} \end{array} \right]^T$ is the steering vector of source l. We then seek to find the spatial frequencies $\{\omega_l\}$ and the number of sources L. The spatial frequencies are related to the DOAs as $\omega_l = \Omega_l d \sin\theta_l / c$ with Ω_l being the center frequency of the lth source, θ_l its DOA, d the sensor spacing, and c the propagation velocity. Assuming that the spatial frequencies are distinct, the columns of $\mathbf{A}$ are linearly independent. The matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times K}$ with $M \leq K$ is the so-called measurement or sensing matrix of compressed sensing (see, e.g., [4, 5]), which here operates across the array exploiting spatial sparsity. We will return to the matter of how to choose M later. This matrix is constructed as a realization of a random process but is assumed known and constant over the N snapshots.

In this paper, we present a subspace-based approach for determining the direction-of-arrivals as well as the number of sources, i.e., the model order. The method is based on a modified covariance matrix model that takes the presence of compressed sensing into account. At this point, it should be stressed that we are not here arguing for the relevance of using compressed sensing in this context, but rather investigating how the associated problems can be solved in a consistent manner (for some applications of compressed sensing, we refer the reader to [5]). The proposed method is based on the concept of angles between subspaces (see, e.g., [9, 10]), which has recently been shown to be applicable to the problem of model order estimation [11].

The remainder of this paper is organized as follows: In Section 2 we develop the covariance matrix for signals of the form (G.1) and discuss the implications of using compressed sensing on the model. We then proceed to present the proposed joint DOA and order estimator in Section 3 and present some results in Section 4. Finally, we conclude our work in Section 5.

2 Modified Covariance Matrix Model

We will now proceed to derive the modified covariance matrix for the compressed sensing scenario. The $M \times M$ covariance matrix of the observed signal is then

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{y}(n)\mathbf{y}^H(n)\right\} = \mathbf{\Phi}\mathbf{A}\mathbf{P}\mathbf{A}^H\mathbf{\Phi}^T + \mathbf{\Phi}\mathbf{Q}\mathbf{\Phi}^T. \tag{G.2}$$

Assuming that the signals of the individual sources in the vector $\mathbf{x}(n)$ are independent and zero-mean, the matrix $\mathbf{P}$ is diagonal and contains the expected power of the individual sources.

From (G.2), it can be observed that the measurement matrix generally changes the covariance matrix of the noise, rendering even white noise colored, and this must be addressed before we proceed. We will here do this by introducing a pre-whitener that takes the presence of colored noise and compressed sensing into account as follows. Let

$\mathbf{C}$ be the Cholesky factor such that

$$\left(\mathbf{\Phi Q \Phi}^T\right)^{-1} = \mathbf{C}^H\mathbf{C}, \tag{G.3}$$

with $\mathbf{C}$ being a square, upper triangular matrix. We note that $\mathbf{\Phi Q \Phi}^T$ is square, positive definite and has full rank and thus invertible with probability close to one. Note that the Cholesky factor is generally complex due to $\mathbf{Q}$ being complex. Then, by multiplying $\mathbf{y}(n)$ by $\mathbf{C}$, we obtain a pre-whitened signal whose covariance matrix is given by

$$\widetilde{\mathbf{R}} = \mathbf{C\Phi A P A}^H\mathbf{\Phi}^T\mathbf{C}^H + \mathbf{I}. \tag{G.4}$$

Note that the noise variance can vary and be unknown without affecting the derivations that follow. Let $\mathbf{U\Lambda U}^H$ be the eigenvalue decomposition (EVD) of $\widetilde{\mathbf{R}}$. Then, $\mathbf{U}$ contains the M orthonormal eigenvectors $\mathbf{u}_m$, and $\mathbf{\Lambda}$ is a diagonal matrix containing the corresponding eigenvalues, λ_m, with $\lambda_1 \geq \ldots \geq \lambda_M$. Let $\mathbf{S}$ be formed from the eigenvectors corresponding to the L most significant eigenvalues, the range $\mathcal{R}(\cdot)$ of which we refer to as the signal subspace. Similarly, let $\mathbf{G}$ be formed from the eigenvectors corresponding to the $M-L$ least significant eigenvalues, i.e., $\mathbf{G} = \begin{bmatrix} \mathbf{u}_{L+1} & \cdots & \mathbf{u}_M \end{bmatrix}$, and $\mathcal{R}(\mathbf{G})$ is referred to as the noise subspace. It can then be shown that the columns of $\mathbf{C\Phi A}$ span the same space as the columns of $\mathbf{S}$, and that $\mathbf{C\Phi A}$ therefore also must be orthogonal to $\mathbf{G}$, i.e.,

$$\mathbf{A}^H\mathbf{\Phi}^T\mathbf{C}^H\mathbf{G} \triangleq \mathbf{A}^H\mathbf{\Xi} = \mathbf{0}, \tag{G.5}$$

or, equivalently, $\mathbf{A}$ should be orthogonal to $\mathbf{\Xi}$. This result is an extension of the basic result used in the MUSIC algorithm as originally proposed in [1], here modified to account for compressed sensing. In practice, $\mathbf{G}$ is of course unknown and an estimate can be obtained from the EVD of the sample covariance matrix. From the above, it can then be seen that to estimate $\mathbf{G}$, we must require that $M > L$. Additionally, we observe that it is required that $\text{rank}(\mathbf{\Phi A}) = L$, which essentially means that we must construct a measurement matrix $\mathbf{\Phi}$ that, regardless of what the spatial frequencies $\{\omega_l\}$ are, must have rows that capture or are likely to capture the column space of $\mathbf{A}$, i.e., $\mathcal{R}(\mathbf{A}) \in \mathcal{R}\left(\mathbf{\Phi}^T\right)$. This is essentially also what the so-called *restricted isometry property* (RIP) of compressed sensing says, and some ways of constructing matrices (and choosing M) that obey this have been proposed in the literature [12]. Until recently [13], these conditions were only shown to hold for what has been referred to as incoherent dictionaries, but such conditions are in direct contradiction with the physics of the considered problem as the individual spatial angles can occur on a continuum of values, corresponding to a highly coherent dictionary. The question still remains, however, how to choose M when L is unknown. We here propose to simply put an upper bound on it and choose M accordingly (see, e.g., [5, 13] for details), and hence facilitate estimating the exact L number of sources as long as it is lower than this bound. Regarding the

number of snapshots N, it must be at least as high as the number of sources, L, to allow identification of the signal subspace and its orthogonal complement, i.e., $N \geq L$. Note that the covariance matrix need not be full rank for this approach to work.

On a related note, the issue of the coherence of the dictionary is also closely related to an advantage that the proposed method holds over methods like basis pursuit [14], since such methods are inherently restricted to finite dictionaries, while the proposed method can (finite precision effects aside) find underlying continuous parameters, corresponding to an infinite and highly coherent dictionary.

3 Measuring Orthogonality

The question is now how to measure the orthogonality between the two matrices $\mathbf{A}$ and $\boldsymbol{\Xi}$. In answering this question, we will turn to the notion of angles between subspaces in linear algebra. Let $\boldsymbol{\Pi}_{\boldsymbol{\Xi}}$ be the projection matrix for the subspace $\mathcal{R}(\boldsymbol{\Xi})$ and $\boldsymbol{\Pi}_A$ the projection matrix for the subspace $\mathcal{R}(\mathbf{A})$. The principal angles between the two subspaces are defined recursively as (see, e.g., [9])

$$\cos(\theta_k) = \max_{\mathbf{y}\in\mathbb{C}^M} \max_{\mathbf{z}\in\mathbb{C}^M} \frac{\mathbf{y}^H \boldsymbol{\Pi}_A \boldsymbol{\Pi}_{\boldsymbol{\Xi}} \mathbf{z}}{\|\mathbf{y}\|_2 \|\mathbf{z}\|_2} \tag{G.6}$$

$$\triangleq \mathbf{y}_k^H \boldsymbol{\Pi}_A \boldsymbol{\Pi}_{\boldsymbol{\Xi}} \mathbf{z}_k = \xi_k, \tag{G.7}$$

for $k = 1, \ldots, \kappa$ and orthogonal vectors $\mathbf{y}^H \mathbf{y}_i = 0$ and $\mathbf{z}^H \mathbf{z}_i = 0$ for $i = 1, \ldots, k-1$. Furthermore, κ is the minimal dimension of the two subspaces, i.e., $\kappa = \min\{L, M-L\}$. As can be seen, $\{\xi_k\}$ are the singular values of the matrix product $\boldsymbol{\Pi}_A \boldsymbol{\Pi}_{\boldsymbol{\Xi}}$. As was shown in [11], a convenient and accurate scalar measure of the angles is the average over cosine to the angles squared, i.e.,

$$\frac{1}{\kappa} \sum_{k=1}^{\kappa} \cos^2(\theta_k) = \frac{1}{\kappa} \|\boldsymbol{\Pi}_A \boldsymbol{\Pi}_{\boldsymbol{\Xi}}\|_F^2. \tag{G.8}$$

For computing this, we need to determine the two projection matrices. For $\boldsymbol{\Xi}$ this is not problem as it has to be calculated only once for each set of snapshots and candidate L. It is, however, more problematic for $\mathbf{A}$, as it depends on all the DOAs—this is also the reason that $\boldsymbol{\Phi}$ and $\mathbf{C}$ are multiplied onto $\mathbf{G}$ rather than $\mathbf{A}$. Noting that the columns in $\mathbf{A}$ are asymptotically orthogonal, i.e., $\lim_{K\to\infty} K \boldsymbol{\Pi}_A = \lim_{K\to\infty} K \mathbf{A} \left(\mathbf{A}^H \mathbf{A}\right)^{-1} \mathbf{A}^H = \mathbf{A}\mathbf{A}^H$, we see that the projection matrix can be simplified significantly. Let $\boldsymbol{\Gamma}$ be an orthogonal basis for $\mathcal{R}(\boldsymbol{\Xi})$. Then its projection matrix is given by $\boldsymbol{\Pi}_{\boldsymbol{\Xi}} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^H$ and (G.8) can be expressed as

$$\frac{1}{\kappa} \sum_{k=1}^{\kappa} \cos^2(\theta_k) = \frac{1}{\kappa} \|\mathbf{A}^H \boldsymbol{\Gamma}\|_F^2. \tag{G.9}$$

This measure can now be used to determine the model order L as well as the spatial frequencies $\{\omega_l\}$ as the parameters that combine to minimize the average angle between the two subspaces, i.e., (see [11] for details)

$$\left(\hat{L}, \{\hat{\omega}_l\}\right) = \arg\min_{L} \frac{1}{\kappa} \sum_{l=1}^{L} \min_{\omega_l} \|\mathbf{a}^H(\omega_l)\mathbf{\Gamma}\|_F^2, \tag{G.10}$$

which follows from the additivity of the Frobenius norm over the columns of $\mathbf{A}$. Note that both $\mathbf{A}$ and $\mathbf{\Gamma}$ depend on L while only $\mathbf{A}$ depends on $\{\omega_l\}$. Hence, (G.10) is *not* equivalent to simply estimating the order by identifying peaks in the pseudo-spectrum. (G.10) is a practical measure as the minimization over $\{\omega_l\}$ is decoupled into L minimizations over one nonlinear parameter. Moreover, the inner products involved in the computation of $\mathbf{a}^H(\omega_l)\mathbf{\Gamma}$ can be efficiently computed using FFTs, or, alternatively, using standard polynomial rooting methods. We note that it can be seen from the definition of angles between subspaces that the measure used in the original MUSIC algorithm is only correct when both the involved matrices consist of orthogonal columns.

4 Results

We will now report some experimental results in the form of root mean square estimation errors (RMSEs) for the spatial frequencies as well as the percentage of correctly estimated model orders. The results were obtained using Monte Carlo simulations with 100 runs for each data point. The dependencies of the RMSE on various factors have been investigated by varying the signal-to-noise ratio (SNR), the number of measurements M retained in the compressed sensing, and the number of snapshots N in separate experiments. As reference, the Cramér-Rao lower bound (CRLB) is reported as well, and we compare to another method capable of joint DOA and order estimation, namely ESPRIT [2] extended to order estimation as proposed in [15]. Measurement matrices generated as realizations of a Gaussian i.i.d. process with $M = 4L$ as has been reported to work well in practice [5] are used, except in the experiment in which M is varied. These were randomized in each Monte Carlo run and were then also compared to the performance without compressed sensing. That is, by setting the measurement matrix equal to the identity matrix, the proposed method reduces to MUSIC, except that it also determines the model order. In the figures to follow, we refer to the case with compressed sensing as CS and the proposed method based on angles between subspaces as AbS. For all the experiments, additive, white, Gaussian noise was used along with $L = 5$ narrowband sources, impinging on the array from different angles. For simplicity in the experiment (and for retaining the same SNR for all sources), these sources were generated having distinct spatial frequencies $\{\omega_l\}$, namely 0.7966, 2.2467, 3.1414, 4.4963, and 6.2727, identical power and i.i.d. uniformly distributed phases. Except when

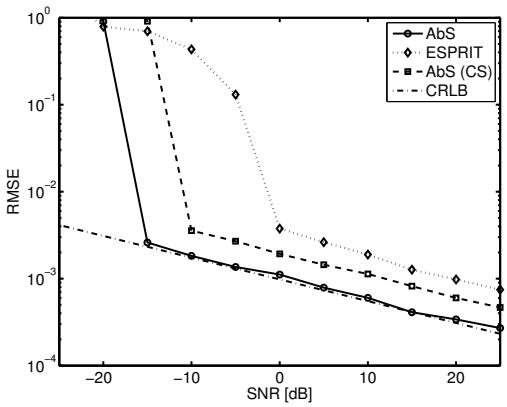

Fig. G.1: RMSE as a function of the SNR.

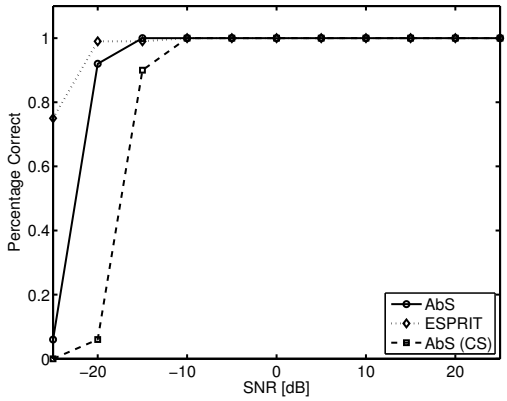

Fig. G.2: Percentage of correctly estimated model orders as a function of the SNR.

otherwise stated, an SNR of 10 dB was used, along with $K = 50$ and $N = 50$. In Figures G.1-G.6, the RMSE and the percentage of correctly estimated model orders are shown as functions of the various parameters. First of all, it can be observed that the MUSIC method performs close to the CRLB when compressed sensing is not used. This is, however, not the case when compressed sensing is used, as a gap can be observed. Interestingly, we observe that this gap is approximately equal to the ratio between K and M, i.e., the performance obtained with compressed sensing is identical to what one would have obtained by simply using M sensors instead of K. It can also be observed that the proposed method determines the correct model order, except under adverse conditions with very low SNRs and low M, and, as M is increased, its performance

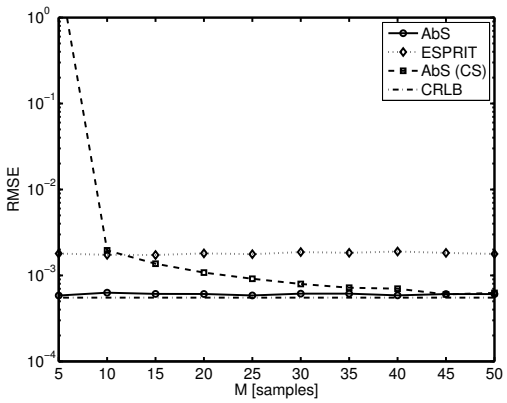

Fig. G.3: RMSE as a function of measurements M.

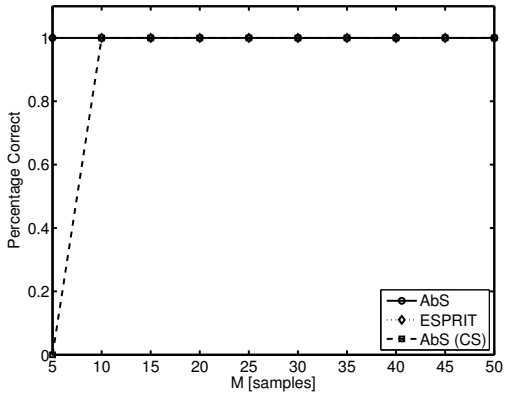

Fig. G.4: Percentage of correctly estimated model orders as a function of the number of measurements M.

approaches the CRLB. Note that several of the methods exhibit identical performance over some intervals in Figures G.4 and G.6, for which reason the curves fall on top of each other. An interesting observation from these experiments is that the threshold behavior changes with the use of compressed sensing, meaning that the estimator breaks down earlier than without compressed sensing. Another observation is that even with compressed sensing, the proposed method outperforms the ESPRIT algorithm in terms of RMSE.

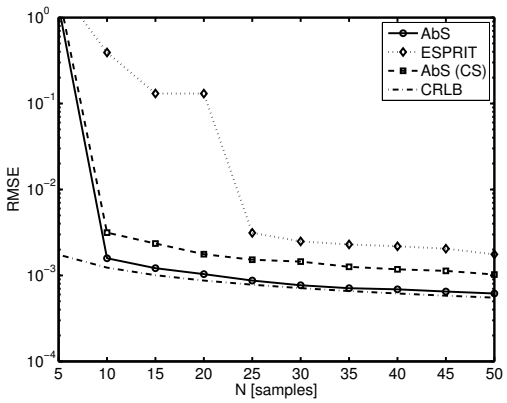

Fig. G.5: RMSE as a function of the number of snapshots N.

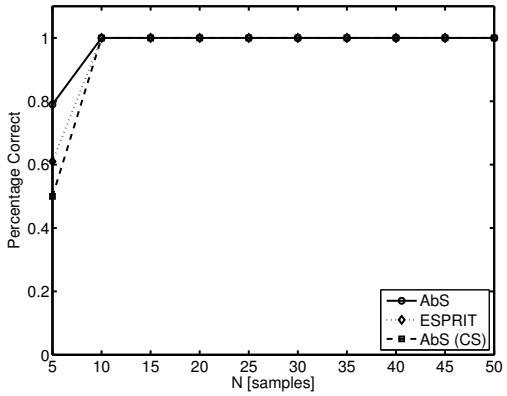

Fig. G.6: Percentage of correctly estimated model orders as a function of the number of the snapshots N.

5 Conclusion

In this paper, we have considered the problem of jointly determining the direction-of-arrivals and the number of sources jointly in arrays employing compressed sensing. We have shown how this problem can be solved using a novel subspace approach based on angles between subspaces. The method was demonstrated to have good performance, resulting in both accurate estimates of the spatial frequencies and the number of sources. Moreover, despite the introduction of compressed sensing it is possible to implement the algorithm using FFTs or root methods provided that a re-orthogonalization step is

introduced. Interestingly, the results show, that the threshold behavior of the method changes with the use of compressed sensing, with the method still outperforming the ESPRIT algorithm.

References

[1] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34(3), pp. 276–280, Mar. 1986.

[2] R. Roy and T. Kailath, "ESPRIT – estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37(7), Jul. 1989.

[3] G. Bienvenu, "Influence of the spatial coherence of the background noise on high resolution passive methods," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, pp. 306–309.

[4] D. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52(4), pp. 1289–1306, apr 2006.

[5] E. Candés and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25(2), pp. 21–30, Mar. 2008.

[6] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53(8), pp. 3010–3022, Aug. 2005.

[7] A. C. Gübüz, J. H. McClellan, and V. Cevher, "A compressive beamforming method," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2008, pp. 2617–2620.

[8] V. Cevher, M. F. Duarte, and R. G. Baraniuk, "Distributed target localization via spatial sparsity," in *Proc. European Signal Processing Conf.*, 2008.

[9] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. The Johns Hopkins University Press, 1996.

[10] R. T. Behrens and L. L. Scharf, "Signal processing applications of oblique projection operators," *IEEE Trans. Signal Process.*, vol. 42, no. 6, pp. 1413–1424, 1994.

[11] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Sinusoidal order estimation using angles between subspaces," *EURASIP J. on Advances in Signal Processing*, pp. 1–11, 2009.

[12] R. Baraniuk, M. Davenport, R. Devore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constr. Approx*, vol. 28(3), pp. 253–263, 2007.

[13] E. J. Candés, Y. C. Eldar, and D. Needell, "Compressed sensing with coherent and redundant dictionaries," May 2010, unpublished manuscript.

[14] S. C., D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1996.

[15] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 450–458, Feb. 2006.